Computationally Engineered CRISPR-SpyCas9 High-Fidelity Variants with Improved Specificity and Reduced Non-specific DNA Damage

Roy Rabinowitz[1,2,*], Oded Shor[2,3,*], Johanna Zerbib[1,*], Shay Herman[1,2], Natalie Zelikson[3], Shreyas Mediwale[3,4], Nataly Yom-Tov[1,2], Uri Ben-David[1], Felix Benninger[2,3,5,#], Daniel Offen[1,2,#]

1. Department of Human Molecular Genetics and Biochemistry, Faculty of Medicine, Tel Aviv University, Israel.
2. Felsenstein Medical Research Center, Petah Tikva, Israel.
3. Faculty of Medicine, Tel Aviv University, Israel.
4. Rina Zaizov Pediatric Hematology Oncology Division, Schneider Children's Medical Center of Israel, Petah Tikvah, Israel.
5. Department of Neurology, Rabin Medical Center, Beilinson Hospital, Petah Tikva, Israel.

*Equally contributing first authors

#Equally contributing last authors

Correspondence:

Daniel Offen (doffen@post.tau.ac.il), Felix Benninger (benninger@tauex.tau.ac.il)

ABSTRACT

The CRISPR-Cas system holds great promise in the treatment of diseases caused by genetic variations. As wildtype SpyCas9 is known to generate many off-target effects, its use in the clinic remains an elusive goal. To date, several high-fidelity Cas9 variants with improved specificity over wildtype SpyCas9 have been described. Each of the variants was engineered either by rational design or experimental directed evolution. However, the potential of improved specificity using both methods is constrained to specific residues due to the requirement of a priori knowledge and challenges in selecting engineered proteins, which arise from factors such as selection pressure, assay limitations, host organism compatibility, and library size and diversity. Therefore, SpyCas9 modifications through in-silico protein engineering allow an efficient specificity improvement while overcoming those limitations. Nevertheless, the lack of living models to test the evolved protein remains a major challenge of computational protein engineering methods. We recently demonstrated the advantage of normal mode analysis to simulate and predict the enzymatic function of SpyCas9 in the presence of mismatches. Here, we report ComPE, a novel computational protein engineering method to modify the protein and measure the vibrational entropy of wildtype or variant SpyCas9 in complex with its sgRNA and target DNA. Using this platform, we discovered novel high-fidelity Cas9 variants with improved specificity. We functionally validated the improved specificity of four variants, and the intact on-target activity in one of them. Lastly, we demonstrate their reduced off-target editing and non-specific gRNA-independent DNA damage, highlighting their advantages for clinical applications. The described method could be applied to a wide range of proteins, from CRISPR-Cas orthologs to distinct proteins in any field where engineered proteins can improve biological processes.

INTRODUCTION

The clustered regulatory interspaced short palindromic repeats (CRISPR)-Cas prokaryotic adaptive immune system[1] was transformed into a powerful gene editing tool, allowing precise, efficient, and simple-to-design genome engineering[2,3]. Directed by a guide RNA (gRNA) molecule, the Cas effector protein binds the target sequence, subsequently leading to its cleavage or other editing alternatives[4–6]. The specificity of the CRISPR-Cas system has a clear impact on both genome-wide off-targeting and allele-specific targeting. Although being considered accurate, the CRISPR-Cas system has its limitations in mismatches discrimination[7–9].

Previously reported engineered *Streptococcus pyogenes* Cas9 (SpyCas9) have been shown to have improved specificity. Each of these high-fidelity (HF) variants was engineered either by rational design[10–15] or experimental directed evolution[16–21]. Rational design is limited to specific residues based on a priori knowledge. Furthermore, in most cases, it is constrained to alanine substitutions. Thus, the guiding principle is neutralizing residues and disruption of the protein activity in selected regions.

In contrast, directed evolution allows the freedom of random mutagenesis and evolvement of the protein by an iterative process of diversification and selection[22]. However, in addition to its time and labor-intensive nature, experimental directed evolution is commonly focused on specific regions of the protein or fails to cultivate potential candidates due to limitations of the host (i.e., bacterial or mammalian cells) and assay. Noteworthy, in many cases, substitution of residues distant from the active site of the protein, unexpectedly leads to significant changes in the protein's activity[23]. Thus, constrained site-directed approaches will most probably overlook such impactful variants.

Comparable to processes that were transformed by digitalization, computational protein engineering could reduce time, labor and costs of research to become a powerful method. The challenge of transforming assays such as directed evolution or deep mutational scanning into a computational process, is the ability to assess the function of the evolving protein in the absence of living cells. Recently, researchers utilized machine learning to improve the on-target activity levels of KKH-*Staphylococcus aureus* Cas9 (SaCas9)[24]. Other methods were developed to improve the stability of proteins based on the energy differences between the reference protein and the variant[25,26]. Such tools were utilized to improve the activity levels of SpyCas9. However, together with the increased activity, the engineered variant demonstrated significant reduction in fidelity leading to elevated levels of off-targeting[27]. To date, no HF Cas variant has been engineered by a computational approach, plausibly due to the tradeoff between activity and specificity[28]. While the tools used to design the new variant aim to improve stability, it is likely that the enhanced stability of Cas9 will lead to compromised specificity. We hypothesized that the stability should be moderately reduced to improve the specificity

of CRISPR-Cas proteins. Thus, to computationally design HF variants, it is necessary to employ a dynamic approach, capable of either increasing or decreasing the stability of the protein, in different circumstances and conditions, and simulating its functionality.

Recently, we demonstrated the ability to predict and simulate the activity of SpyCas9 in the presence of mismatches using normal mode analysis (NMA)[29]. By calculating the vibrational entropy (S_{vib}) of the SpyCas9 protein in complex with its single guide RNA (sgRNA) and target DNA, we obtained strong correlations between the NMA scores and the empirical SpyCas9 specificity profile[8]. Furthermore, we also demonstrated the compatibility of our approach for modified proteins, as we found a strong correlation between modelled known HF variants and their NMA scores[29]. In addition, NMA provides improvement in resolution compared to other methods to model structure-function relations, even when using coarse-grained NMA to overcome the computational limiting factor of analyzing large numbers of atoms in a protein[30,31]. Based on this success, we hypothesized that NMA could be utilized for protein engineering to obtain novel HF Cas9 variants by destabilization. In this work, we present a novel approach to protein engineering, termed Computational Protein Engineering (ComPE) mediated by NMA, which relies on S_{vib} changes derived from a protein's structure. Utilizing ComPE, we evolve SpyCas9 and pinpoint specific mutations that result in destabilization. We characterize nine candidate variants predicted by our ComPE method and experimentally validate their specificity attributes. Lastly, we demonstrate that HF Cas9 variants reduce gRNA-independent non-specific DNA damage, further enhancing the benefits of HF engineered variants in clinical applications.

RESULTS

NMA as the basis for the ComPE platform

NMA is a dynamic approach to study the function of proteins. We previously reported the use of NMA to link between genotype and phenotype in context of disease-causing mutations[32–35], and its use to assess the specificity profile of SpyCas9 in the presence of mismatches[29]. Inspired by its ability to simulate the protein function, we used NMA to build ComPE, an in-silico method combining principles of directed evolution and deep mutational scanning (**Fig. 1**). While experimental methods rely on survival or reporter-based assays, we first collected available data of engineered variants to establish the vibrational entropy (S_{vib}) patterns in relation to experimental empirical data. The diversification stage has several parameters: fully or partially random, single or multiple substitutions, avoidance of active site mutagenesis, substitution to all or limited amino acids and more mutagenesis parameters. This stage goes from altering the primary structure of the protein to modeling the allosteric changes in the quaternary structure. Finally, we assessed the S_{vib} profile of each new generated structure using

NMA and compared it to the generated S_{vib} profiles of known variants to identify potential improved candidates.

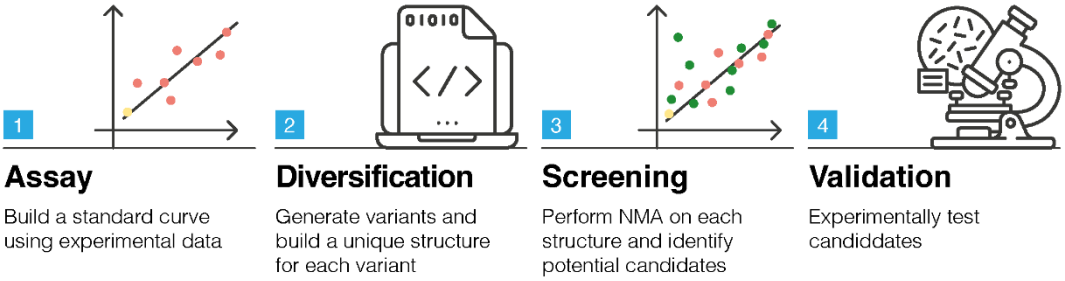


Figure 1: General scheme of the ComPE pipeline.

The Assay: Vibrational entropy of HF variants correlate with the activity and specificity

The association between the structure of a protein and its function is the underlying foundation of extrapolating the properties of the evolving protein. It is commonly accepted to classify Cas9 functions into on-target activity and specificity. While both parameters derive from the protein's activity level, there is a clear negative correlation between the two. For example, HF variants with extremely high specificity (e.g., evoCas9), suffer from a severe reduction in on-target activity levels[20]. Since most engineered Cas9 HF variants exhibit impaired on-target activity, we aimed to discover novel variants with improved specificity and intact on-target activity. We used NMA to calculate the S_{vib} of SpyCas9 and six HF variants (eSpCas9(1.1), SpCas9-HF1, HiFi-Cas9, HypaCas9, evoCas9 and Sniper-Cas9) and compared their S_{vib} values to previously reported experimental measured activity and specificity[20]. As the correlation between the S_{vib} and the empirical data increases, the accuracy of our predictions improves. Therefore, we sought to assess the S_{vib} of which residue yields the best correlation. Once a structure was generated for each HF variant (based on the structure of WT SpyCas9, PDB:5F9R), we examined which residue returns the highest correlation coefficient (r), for all seven Cas9 enzymes, between the S_{vib} and the activity or specificity scores (**Fig. 2**). The selected residue for S_{vib} calculation is not implied to be a candidate for mutagenesis, but rather a reliable indicator of the S_{vib} and protein function correlation. We found that H641 and K1107 had the highest correlation with the activity (r=-0.9885, Fig. 2A-B) and specificity (r=0.7402, Fig. 2C-D), respectively. These residues were used for our further analyses in this study. The contrary direction of the activity and specificity correlations implies the tradeoff between the two.

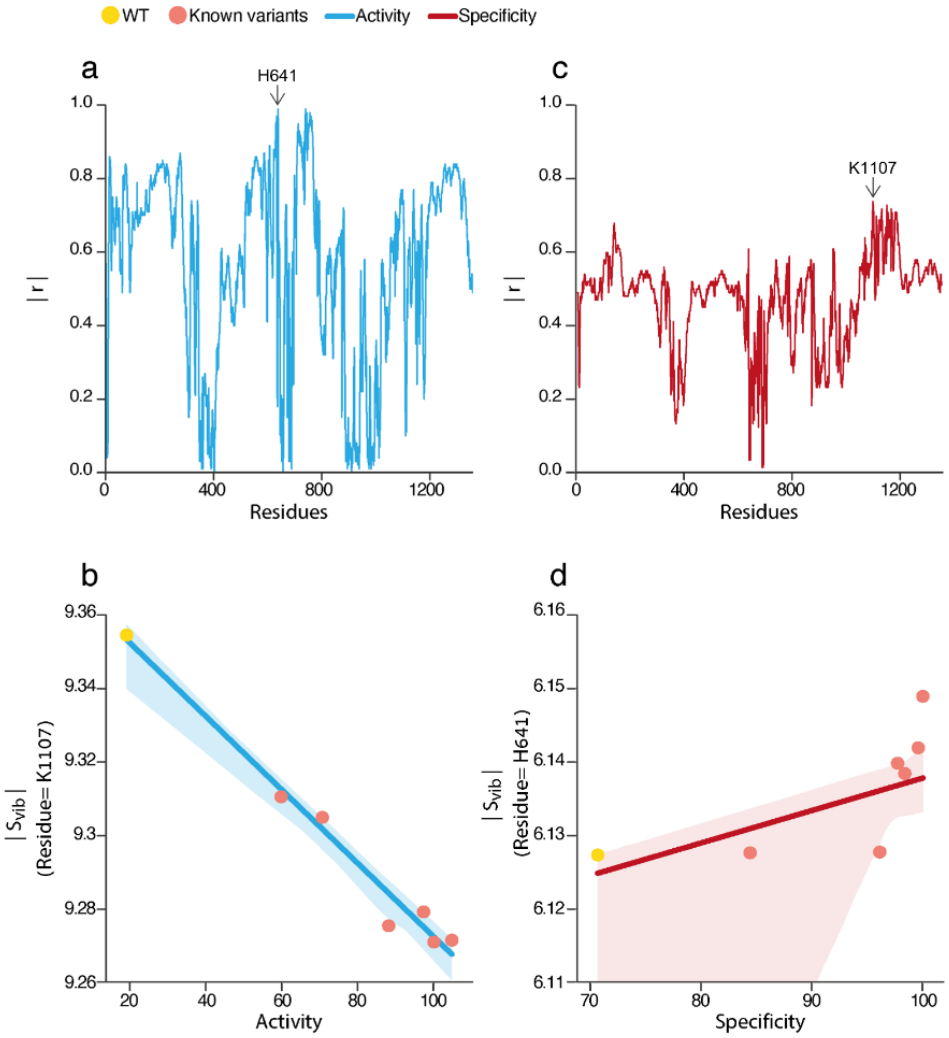


Figure 2: Vibrational entropy values correlate with empirical activity and specificity. **a,** Activity-NMA correlation coefficient |r| was calculated for each of the residues of all seven proteins (WT + 6 HF variants) to assess which residue provides the best correlation for all proteins. **b,** The Activity-NMA correlation plot describes the experimental activity values for each variant and the calculated S_{vib} absolute values as measured at residue H641. The Pearson correlation coefficient is $r = -0.9885$. $P\ value = 3x10^{-5}$. **c,** Specificity-NMA correlation coefficient |r| was calculated for each of the residues of all seven proteins to assess which residue provides the best correlation for all proteins. **d,** The specificity-NMA correlation plot describes the experimental activity values for each variant and the calculated S_{vib} absolute values as measured at residue K1107. The Pearson correlation coefficient is $r = 0.7402$. $P\ value = 0.0571$. All correlation plots are shown with a 95% confidence interval.

Diversification and Screening: In-silico deep mutational scanning and selection of HF Cas9 candidates

To discover novel single-mutation SpyCas9 variants with improved specificity, we diversified its structure by substituting each residue of the protein (1,362 residues) with every other amino acid (19 alternatives). In total, we generated and analyzed 25,878 new structures by NMA to calculate their S_{vib}. The S_{vib} values of each variant were reported on an empirical cumulative distribution function

(eCDF) graph. As expected, we observed that most single-mutation substitutions merely altered the S_{vib} of the protein, having most of the variants concentrated around $Median_{(Activity)} = -2.213$ (**Fig. 3A**) and $Median_{(Specificity)} = -2.033$ (**Fig. 3B**). Interestingly, previously reported HF variants have shown increased S_{vib} compared to WT SpyCas9 for both the activity and specificity, excluding Sniper-Cas9 showing lower S_{vib} on the specificity plot. Taken together, these results strengthen our hypothesis that moderate destabilization of Cas9, depicted with mild increase of S_{vib}, would lead to improved specificity of the protein. We selected nine candidates from the top 0.25% ($F_{(entropy)} \geq 0.9975$), all characterized with higher S_{vib} than known HF variants and mutated in different regions and domains of the protein. The selected candidates were not previously described in other previously reported HF variants: S512R, T519V, N692V, K735W, Q739R, V743N, K1129T, K1231I and R1359Q (in a sequence-chronological order). Interestingly, the S_{vib} of WT SpyCas9 was found to have $F_{(Svib)} = \sim 0.5$ for both activity and specificity, meaning that about half of the single amino acid substitution variants have lower S_{vib} compared to WT SpyCas9, while the other half have higher S_{vib}. The nature of the WT enzyme is to be stable and specific on one hand, but also adjustable to small changes on the other hand. Therefore, this observation may imply an evolutionary balance, leading to stability suitable for the function of the enzyme, and favorable over other small changes in the sequence of the protein. It is noteworthy that HypaCas9, which was generated by rational design, contains the N692A substitution along with three other alanine substitutions[12]. Our independent observation of the N692V candidate strengthens our hypothesis that increased S_{vib} is the underlying mechanism of increased specificity in HF Cas9 variants.

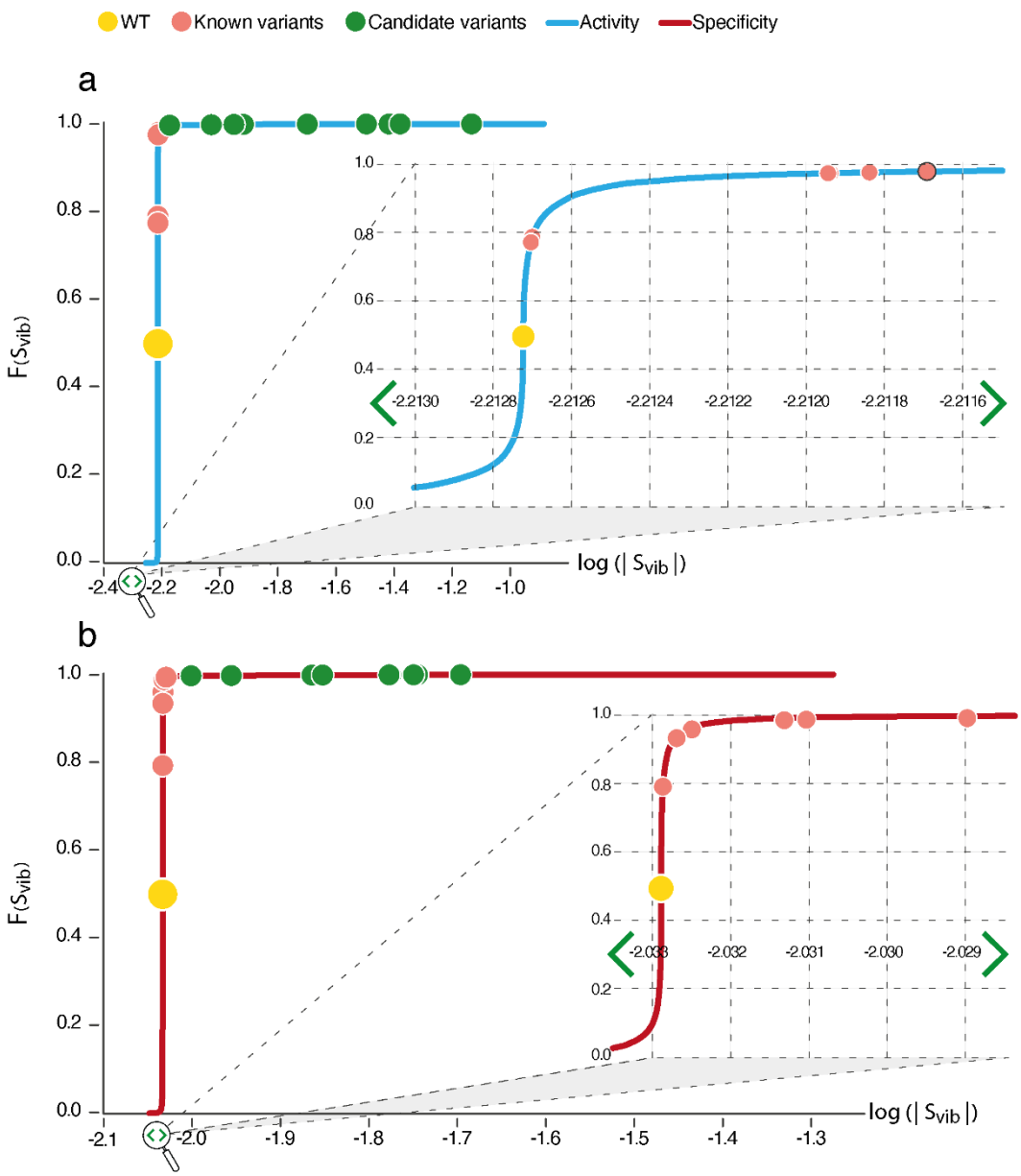


Figure 3: In-silico deep mutational scanning of SpyCas9 variants and analyses of their activity and specificity.
a, eCDF plot of the proportion of variants and their log ($|S_{vib}|$) values, based on the S_{vib} score of residue H641 to assess the activity of all possible single-amino acid substitutions. The Y axis, $F(S_{vib})$, describes the proportion of variants with a given value of log ($|S_{vib}|$) or lower. The blue line represents the deep mutational scanning derivative variants (25,878 variants), while the dots represent WT (yellow) or the known HF Cas9 enzymes (light red) and the chosen candidates (green). The magnification box displays the sigmoid area of the curve. **b,** eCDF plot of the proportion of variants and their log ($|S_{vib}|$) values, based on the S_{vib} score of residue K1107 to assess the specificity of all possible single-amino acid substitutions. The red line represents the deep mutational scanning derivative variants.

HF Cas9 candidates have reduced genome-wide off-targets

We then aimed to validate that the selected destabilized variants show increased specificity while conserving high activity of the enzyme. We generated a plasmid library of these nine candidates, as

well as WT SpyCas9 and previously reported HF variants eSpCas9(1.1), HiFi-Cas9 and Sniper-Cas9, into the same backbone to eliminate differences derived from different plasmid elements (i.e., regulatory elements and codon usage). We assessed the genome-wide off-targeting activity of our candidates compared to WT SpyCas9 and the known HF variants using a rhAmpSeq assay, a Next Generation Sequencing (NGS)-based method of on- and off-target evaluation using multiplex PCR[36]. We co-transfected into HEK293FT cells any of the Cas plasmids with three sgRNAs targeting different loci with previously reported off-targets (EMX1, VEGFA and AAVS1)[20,36], and extracted the genomic DNA 96 hours post-transfection. We first investigated the on-target activity of our candidate variants compared to WT SpyCas9, as many of the engineered HF variants are known to have reduced levels of activity[20,37]. Indeed, we observed decreased levels of activity in eSpCas9(1.1) and in some of our candidates (**Fig. 4a**). We then analyzed the off-target activity of the candidate variants, and identified that four of them (S512R, N692V, K735W and V743N) have a compelling reduced off-target activity. Moreover, we found reduced protein levels of the same candidates compared to WT and known HF variants, 72 hours following the transfection in HEK293FT cells (**Sup. Fig. 1**), demonstrating the lower stability of these variants. Together, these results highlight four novel HF variants, and the N692V variant in particular, with both good on-target activity and reduced off-targeting.

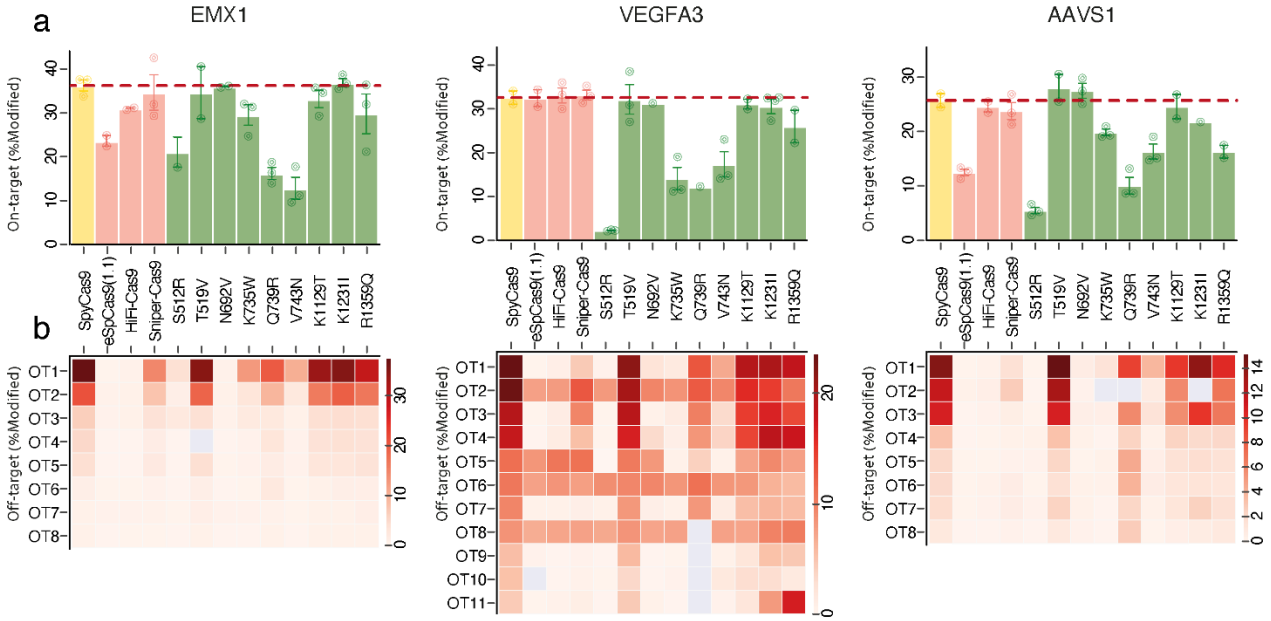


Figure 4: On-target activity and off-target analysis of the HF Cas9 variants.
a, The on-target activity of WT SpyCas9 and the tested HF variants with three sgRNAs targeting EMX1, VEGFA and AAVS1. The activity is measured as the percentage of modified reads. The line indicates the average activity of WT SpyCas9. Error bars describe the standard error of the mean. **b,** Off-targeting by WT SpyCas9 and the tested HF variants with the three sgRNAs. Each off-target (OT) site is represented in a row. The white-red scale represents the percentage of modified reads for each locus. Only OT sites with >1% modified reads are shown. The data in both panels (a,b) represent three biological repeats for most samples. Grey boxes (b) are loci for which there were no reads (in any of the repeats).

Single-base resolution – specificity profile of the HF Cas9 candidates

To further investigate our destabilized Cas9 candidates, we then characterized the specificity profile of each variant. In contrast to genome-wide off-targets analysis, the specificity profile describes the single base resolution of Cas enzymes. Previous studies demonstrated the specificity profile of Cas enzymes using high-throughput methods, generally based on either flow cytometry[9,38,39] or deep sequencing[8,40,41]. It offers a deeper understanding of the mutational effects on the engineered variants, and it may explain the off-targeting patterns observed in unbiased methods such as rhAmpSeq, GUIDE-seq and others. Motivated by our previous off-targets analysis using rhAmpSeq, we employed a GFP disruption assay[9] in EGFP-stable HEK293FT cells, using sgRNAs targeting the GFP gene with mismatches in each of the positions within the gRNA. We co-transfected the cells with both a Cas9 variant and an sgRNA and measured the activity of the Cas9 variant by quantifying the GFP fluorescence intensity using flow cytometry (**Fig. 5A**). We performed 20 different sgRNA combinations for each variant; we conserved one perfectly matched sgRNA to assess the maximal GFP reduction, while the other 19 sgRNAs carry a transversion mismatch (A↔T and G↔C) in positions 1-19. Since the hU6 promoter, which drives the expression of the sgRNA, requires a 5'G to initiate transcription, the 20th position remained unaltered (**Sup. File 1**).

Consistent with the rhAmpSeq results (**Fig. 4**), the four variants S512R, N692V, K735W and V743N exhibited improved sensitivity to mismatches, as demonstrated by the reduced EGFP disruption signals (**Fig. 5B**). We found different patterns of specificity between two groups of variants: the specificity of S512R and N692V was improved all along the gRNA, while K735W and V743N displayed HF properties along the seed region of the gRNA (positions one through eleven). It is noteworthy that residues 512 and 692 are part of the REC III domain (REC lobe), while residues 735 and 743 lay within the RuvC II domain (NUC lobe) of *Spy*Cas9. Mutations of residues at both regions are commonly abundant in the variety of engineered Cas9 variants[7]. Furthermore, when we normalized the disruption created by each variant to WT Cas9, we found that S512R and N692V induced comparable disruption than the previously reported HF variants (**Fig. 5C**). Surprisingly, although being partially HF, K735W also demonstrated high specificity compared to all other tested variants within the positions of the seed sequence. Taken together, these results validate the possibility to computationally obtain high specificity variants, with comparable or better results to the WT SpyCas9 and three previously reported HF Cas9 variants.

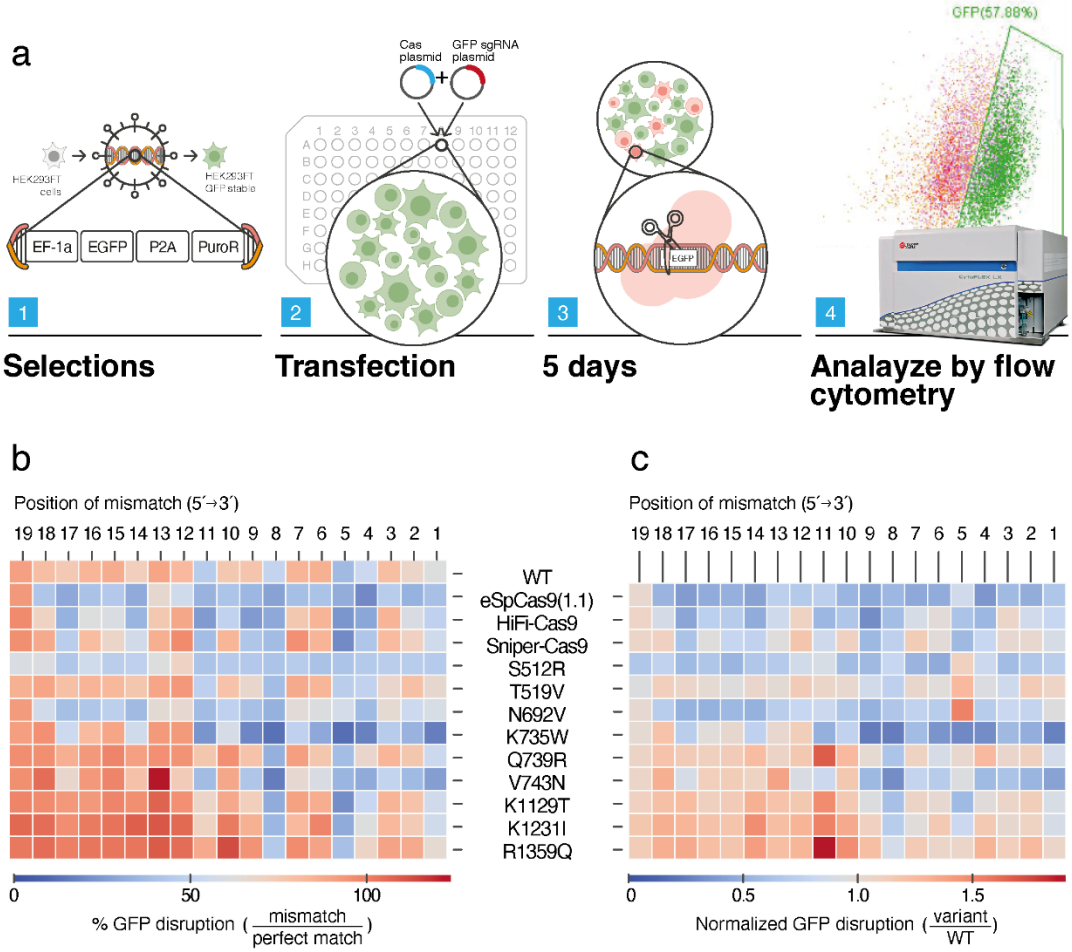


Figure 5: GFP-disruption assay reveals the specificity profile of HF SpyCas9 variants. **a,** Description of the experimental workflow. **b,** Heatmap representation of the GFP disruption by different Cas9 variants combined with mismatched gRNAs analyzed using flow cytometry. All values were calculated as the ratio of GFP disruption in the presence of a mismatch relative to the disruption with a perfect match gRNA. Thus, 100% indicates equal activity of the Cas9 variant combined with both the perfect match gRNA and in the presence of a mismatch. Values below 100% (towards the blue shades) indicate improved sensitivity to a mismatch in each position. Each data point represents the average of two biological replicates. **c,** Normalization of Figure 5c to demonstrate the sensitivity of Cas9 variants compared to WT SpyCas9. Values around 1 indicate no change in mismatch sensitivity compared to WT SpyCas9. Above 1 (dark red shades), sensitivity is impaired, while below 1 (blue shades), mismatch sensitivity is improved compared to WT SpyCas9.

HF Cas9 variants reduce non-specific DNA damage

CRISPR-Cas9 was previously demonstrated to cause p53 activation in cell cultures[42,43]. Although being thought inactive in the absence of a gRNA, the introduction of Cas9 into cells without a gRNA also results in non-specific DNA damage (i.e., gH2AX foci and activation of the P53 pathway)[44]. Kinetic studies of HF Cas9 variants reported that engineered variants act more slowly[28,45], probably due to requirement of increased activation energy. We hypothesized that HF variants, and destabilized variants in particular, are less likely to undergo spontaneous conformational transitions, therefore less likely to activate p53 pathway in comparison to WT SpyCas9. To assess the effect of HF variants on

non-specific DNA damage, we transfected p53-WT MCF7 cells with WT SpyCas9 or with HF variants, and measured the expression of common p53-pathway downstream targets: GADD45, DDB2, RRM2B, XPC, TIGAR, MDM2 and CCGN1. As previously reported[44], transfection of WT SpyCas9 significantly increased the expression of all tested p53 downstream target genes (**Sup. Fig. 2**). Using a p53 pathway activity score derived from the average gene expression changes of the p53 pathways, we found that introduction of both previously reported (eSpCas9(1.1) and HiFi-Cas9) and destabilized (S512R and N692V) Cas9 variants significantly reduced p53 pathway activation compared to WT Cas9 (**Fig. 6A**), consistent with the reduced off-target and specificity profiles of all these variants (**Fig. 4B, Fig. 5D**). Interestingly, our destabilized HF variants, S512R and N692V, also significantly less activated p53 pathway compared to previously reported HF variants (eSpCas9(1.1) and HiFi-Cas9, **Fig. 6B**). Together, these results show that our computationally engineered destabilized Cas9 variants are less susceptible to induce non-specific DNA damage than WT SpyCas9 and other previously reported HF variants. As we also showed that N692V has an intact on-targeting level, our results suggest that this destabilized variant may be of particular advantage for gene editing with minimal cellular damage.

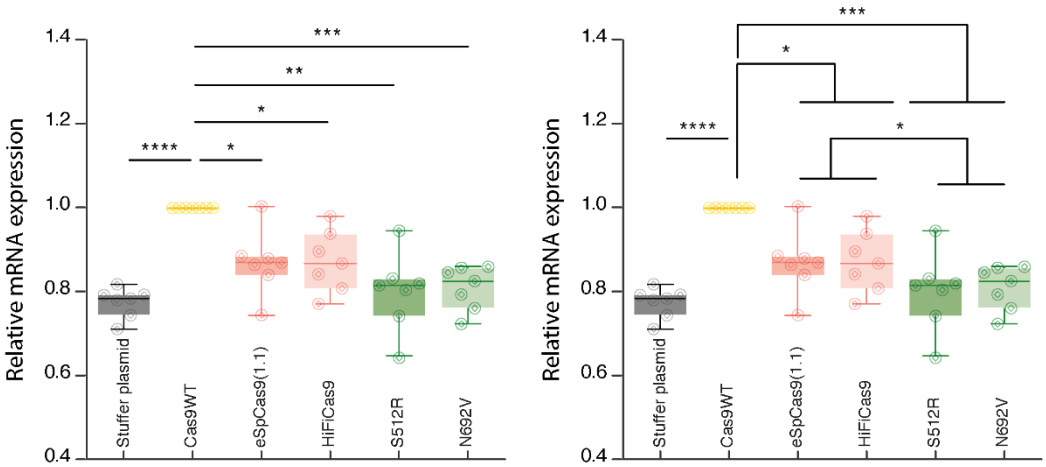


Figure 6: High fidelity variants less activate p53 pathway than WT Cas9.
a, Pooled gene expression results of p53 pathway downstream targets (GADD45, DDB2, RRM2B, XPC, TIGAR, MDM2, and CCGN1) for WT SpyCas9, eSpCas9(1.1), HiFi-Cas9, S512R and N692V. All HF variants significantly less activate p53 pathway than WT SpyCas9. *p=0.0144 and p=0.0127, **p=0.0042, ***p=0.0003, ****p<0.0001 for eSpCas9(1.1), HiFi-Cas9, S512R, N692V and Stuffer respectively; RM ANOVA, Dunnett's multiple comparison. **b,** Group results of pooled gene expression results of p53 downstream targets. Candidate HF variants activate p53 pathway less than commercial HF variants. *p=0.0175 and p=0.0127, ***p=0.0004, ****p<0.0001 for WT/Published, Published/Candidate, WT/Candidate and Stuffer respectively; RM ANOVA, Sidak's multiple comparison.

DISCUSSION

We present ComPE, a novel approach for computational vibrational entropy-based deep mutational scanning. We employed ComPE to engineer SpyCas9 variants with improved specificity. While previously described Cas9 HF variants were engineered either by rational design or experimental directed evolution, here we report for the first time the use of an unbiased in-silico assay to predict and generate engineered variants with improved specificity (**Figs. 1-3**). Four (out of nine) of our predicted candidates were found to have improved specificity by experimental assays, pointing them as successful HF variants (**Figs. 4-5**). Notably, N692V also demonstrated both intact on-target and low off-target activities. As CRISPR-based therapeutics show promising results in clinical trials[46–49] and move forward toward applicable solutions to treat genetic conditions, engineered Cas enzymes have a great potential in increasing the safety of such treatments by reducing unintended off-target activity. Importantly, we also report that the use of HF variants leads to little-to-no non-specific DNA damage (**Fig. 6**). The attenuated activation of the p53 pathway may reduce the selection for p53-inactivating mutations[44], and also reduce undesired on-target effects, such as aneuploidy formation[50,51]. It may also have further benefits in homologous-directed repair (HDR)-based editing, as it is known that p53 inhibits the HDR process[42,43,52]. Previous studies also show that HF variants have increased HDR rates[53–55] . We propose that reduced activation of the p53 pathway may be responsible for HDR enhancement in HF variants editing. This hypothesis needs to be validated in a future study. Taken together, HF variants show great advantages for clinical and research applications.

Previous studies demonstrated that NMA could be utilized to analyze the effect of point mutations on proteins in terms of stability and flexibility[31,56]. Here, we report for the first time, the use of NMA to perform deep mutational scanning, based on reference experimental data, to improve a protein's function in terms of binding specificity to binding counterparts (i.e., DNA, RNA, or other proteins). As binding properties can be described using entropy, we believe that our approach is also relevant for engineering different biological macromolecules, and the interactions between them. It is noteworthy that the mutations we report were non-trivial substitutions between different groups of amino acids with distinct properties (S512R, N692V, K735W and V743N), combining the best of both rational design and experimental directed evolution methods. Moreover, out of the four identified point mutations, two were found to be in close proximity to the gRNA or DNA (N692 and K735), while the other two (S512 and V743) are not in close contact with nucleic acids in the structure (**Sup. Fig. 3**). It is likely that allosteric effects account for the modified function of the protein. This observation emphasizes the impact of our method, similar to random directed evolution, but different from rational design: we identified beneficial mutations remote from the active site or ligand-binding residues. Another advantage of our computational mutagenesis approach is that the mutations are

installed in the protein level (i.e., mutagenesis of amino acids in the structure), instead at the DNA level in experimental directed evolution. Our proposed amino-acid mutations (N↔V, T→V and K→W) required the alteration of two nucleotides in the codon, thus are less likely to spontaneously arise during an experimental directed evolution campaign.

Currently, ComPE is limited by computational resources and inefficient processes. It can be assumed that in the future, advanced capabilities will enable multiple iterations of mutagenesis, enabling computational directed evolution and prediction of more complex combinatorial mutagenesis. Furthermore, another limitation is the requirement of a protein structure. This limitation could have been potentially solved by recent advances in protein structure prediction[57,58]. However, these and other tools are yet to be able to predict the structure of a complex with multiple counterparts (e.g., SpyCas9 bound to its sgRNA and target DNA), and their interactions. Such capabilities, combined with entropy-based methods, would greatly advance the field of computational protein engineering and drug discovery.

METHODS

NMA, entropy calculations and structure alterations

The structure of the SpyCas9 complex was taken from the Protein Data Bank (PDB-101; accession numbers PDB: 5F9R[59]). Next, we performed in silico bases mutagenesis of the gRNA (chain A) and DNA (chain C – TS-DNA and chain D – NTS-DNA) to the gRNA and DNA sequences as previously described[29]. The in silico mutagenesis of nucleic acids was made using X3dna- DSSR (https://x3dna.org/) Linux package[60–63]. The resulting structure was used as our WT structure. Further on, using the mutagenesis plugin in PyMol Molecular Graphics System Version 1.8 (Schrödinger, LLC., Cambridge, MA), we created all possible variant structures with single amino acid residue mutagenesis (total of 25,878 single residue mutation structures). The WT and each of the possible single residue mutated structures were analyzed using ENCoM coarse-grained NMA to evaluate the effect of the mutation on the vibrational entropy of the protein. This method is based on the entropic considerations C package of ENCoM[30] available at the ENCoM development website (https://github.com/NRGlab/ENCoM), compiled and used on a Ubuntu platform (Canonical Group, UK). For each analyzed variant, we calculated the vibrational entropy profile of the structure. The same approach was taken to generate the known HF variants to establish the correlation curves. Finding the most correlative residue vibrational entropy of the six high fidelity structures and their activity or specificity scores was performed using MATLAB software (MathWorks, Natick, MA).

Cell culture, transduction and transfection

HEK293FT cells were grown in DMEM supplemented with 2% L-glu, (Gibco 41965039), 10% Fetal Bovine Serum (Gibco 10270-106) and 1% Penicillin-Streptomycin (Gibco 15070063). MCF7 cells were grown in RPMI (Sartorius) supplemented with 4mM Glutamine (Sartorius), 10% Fetal Bovine Serum (Gibco) and 1% Penicillin-Streptomycin (Sigma). All cells were maintained at 37 °C with 5% CO2.

To establish the EGFP-PEST stable cell line, HEK293FT cells were transduced with a lentivirus carrying the EGFP-PEST-P2A-PuroR construct. Following transduction, infected cells were selected with puromycin (Thermo-Fisher J67236, 1:1000) for two weeks. For Cas9 activity assays (rhAmpSeq and GFP disruption), HEK293FT cells were transfected with 40ng plasmid DNA/well (sgRNA:Cas ratio 1:1, supplemented with mCherry reporter plasmid for transfection assessment and flow cytometry gating), using 0.2ul/well Lipofectamine3000 (Invitrogen L3000015) and OptiMEM (Gibco 31985-047), following manufacturer's protocol. All transfections were carried in 96-well plates using 8,000 cells/well. Transfection media were replaced 24hr post-transfection.

EMX1 and VEGFA site 3 gRNA sequences were cloned into BPK1520[64] backbone according the standard protocol. All other gRNAs and Cas enzymes were cloned into expression vectors by VectorBuilder.

Estimation of p53 pathway activation following introduction of Cas9 variants was performed as previously described[44]. Briefly, MCF7 cells were transfected with 2ug of DNA constructs using Trans*IT*-LT1 reagent (Mirus), following manufacturer's protocol. RNA was extracted 72hrs post transfection.

DNA extraction

For rhAmpSeq analyses DNA was extracted 96hrs post transfection, as previously described[65]. Briefly, cells were carefully washed with PBS and then lysed with 150ul lysis buffer (10mM Tris-HCl pH8, Sigma T3038; 0.05% SDS, Bio-Basic SD8119; 35ug/ml Proteinase K, Roche 03-115-879-001). After 1hr incubation at 37 °C, lysed cells were pipetted incubated at 80°C for 30min using a thermocycler. The solution of extracted DNA was used as the DNA template for the NGS analysis.

rhAmpSeq analysis

For each target (EMX1, VEGFA and AAVS1), a PCR primers panel was designed using IDT's rhAmpSeq designated tool (**supplementary file 1**)[36]. Library preparation and NGS were conducted according to IDT's protocol[36]. CRISPResso2[66] was used to analyze NGS data and assess the on and off-target activity of the Cas variants for each gRNA (using the CRISPRessoPooled utility). Primer panels for EMX1 and VEGFA site 3 were designed according to GUIDE-seq and TTISS data as previously reported[20]. The

AAVS1 panel was designed based on GUIDE-seq data from IDT's rhAmpSeq protocol[36]. The reported off-targets data used in this study were based on experiments conducted under the same conditions (i.e., cell line and gRNA sequence). Each genomic DNA sample was used to assess both on and off target activity.

GFP disruption assay

EGFP-PEST stable cells were co-transfected with one of the sgRNA plasmids and one of the Cas plasmids, together with a reporter plasmid. Five days post transfection cells were trypsinized, incubated 5 min at 37 °C, and neutralized by 150ul FBS-enriched FACS buffer (PBS; 5% FBS; 25mM HEPES Biological industries (Israel) 03-025; 5mM EDTA, Biological industries (Israel) 01-862). The suspended cells were filtered using a mesh-capped plate (Merck MANMN4010) before flow cytometry analysis. Cells were analyzed by flow cytometry (CytoFLEX S, Beckman Coulter) to assess GFP-disruption rates. Decrease in GFP levels was assessed by measuring the GFP positive cells, based on GFP+ untreated cells, out of mCherry positive cells. The GFP disruption rate was calculated as following:

$$\%\,GFP\ disruption\ = 100 \cdot \frac{100 - \%GFP^{+} - \%background}{\%\,perfect\ match\ sgRNA}$$

Each plate contained its internal positive and negative controls for normalization.

Protein extraction and western blot

Transfected cells were harvested 72 hours post transfection and treated with RIPA buffer (#89901, Thermo Fisher Scientific, Waltham, MA) freshly supplemented with protease and phosphatase inhibitors (1:100, #78441, Thermo Fisher Scientific), incubated 30 min on ice, and centrifuged at 14,000 RPM for 20 min at 4ºC. Protein levels were determined using a bicinchoninic acid assay kit (BCA, #A-4503, Thermo Fisher Scientific). Supernatants were stored at −80ºC until used. Thereafter, 30µg of protein was separated by electrophoresis through 7.5% polyacrylamide gels (Mini-PROTEAN, #456-8024, Bio-Rad) and transferred to nitrocellulose membranes (#1704159, Bio-Rad) using Trans-Blot Turbo Transfer System equipment (Bio-Rad). The membranes were blocked using SuperBlock (#37535, Thermo Fisher Scientific) with 0.1% Tween 20 (#P1379-100ML, Thermo Fisher Scientific) and probed overnight at 4ºC with the primary antibodies: CAS9 (1:1000, MAB #146977, A9-3A3 Cell-Signaling). The membranes were washed three times with Tris buffered saline (#1706435, Bio Rad, CA, USA) and 0.1% Tris Buffer Saline Tween 20 (TBST) and incubated for 1 h at room temperature with secondary antibodies: goat anti-mouse or goat anti-rabbit IRDye® 800CW/680CW (1:10,000, Licor,

Lincoln, NE, USA) for one hour at RT. The membranes were then developed with Odyssey Imager CLx (Licor). As a control for protein loading, blots were probed for mouse anti-Glyceraldehyde 3-phosphate dehydrogenase (GAPDH, 1:1000, MAB #2118, Cell-Signaling) using the same procedures. Data were calculated as the ratio of mean target protein intensity to GAPDH intensity. Densitometric analysis of Western blots was performed using Image Studio Lite software (Licor).

RNA extraction and qRT-PCR

To assess p53 pathway activation, transfected MCF7 cells were harvested using Bio-TRI® (BioLabs) and RNA was extracted following manufacturer's protocol. cDNA was amplified using GoScript™ Reverse Transfection System (Promega), using oligoDT primers, following manufacturer's protocol. qRT-PCR was performed using Sybr Green (Invitrogen), and quantification was performed using ΔCT method. All primers used for detection of p53 pathway targets are available in **Sup. File 1**.

Statistical tests

The GraphPad Prism v9.4.1 software was used to produce all charts and analyze data. Statistical tests were performed as described in the figure legends and P values of ≤0.05 are labeled with a single asterisk (*), in contrast to P values of ≤0.01 (**), ≤0.001 (***), or ≤0.0001 (****); where not indicated, P values are non-significant. In all relevant figure panels, values of mean±SEM are reported, and the exact n value is described in each figure legend.

ACKNOWLEDGEMENTS

The authors would like to thank Michal Shor for her assistance with Figures preparation. This work was supported by the Israel Science Foundation (grant #1010/16 to F.B.), a generous donation from Varda and Boaz Dotan (to F.B.), the European Research Council (ERC Starting Grant # 945674 to U.B.-D.) and by the Israel Science Foundation (grant #1339/18 to U.B.-D.). R.R. was supported by external Ph.D. scholarships from the ''Dan David Prize" and "Teva BioInnovation Fellowship" and travel awards from the TAU Constantiner Institute and the Gertner Intitute. J.Z. was supported by a fellowship of the Israeli Ministry for Immigrant Absorption, by the Pfizer-Wexler Excellence Scholarship, and by travel awards from the TAU Constantiner Institute and Cancer Biology Research Center. The BPK1520 plasmid was a kind gift from Keith Joung (Addgene plasmid #65777; http://n2t.net/addgene:65777;RRID:Addgene_65777).

REFERENCES

1. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science (80-.).* **315**, 1709–1712 (2007).

2. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* **337**, 816–21 (2012).

3. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (80-.).* **339**, 819–23 (2013).

4. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

5. Anzalone, A. V *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* (2019). doi:10.1038/s41586-019-1711-4

6. Wang, H. *et al.* CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. *Cell* **175**, 1405-1417.e14 (2018).

7. Rabinowitz, R. & Offen, D. Single-Base Resolution: Increasing the Specificity of the CRISPR-Cas System in Gene Editing. *Mol. Ther.* (2020). doi:10.1016/j.ymthe.2020.11.009

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

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

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

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

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

13. Bratovič, M. *et al.* Bridge helix arginines play a critical role in Cas9 sensitivity to mismatches. *Nat. Chem. Biol.* (2020). doi:10.1038/s41589-020-0490-4

14. Tan, Y. *et al.* Rationally engineered Staphylococcus aureus Cas9 nucleases with high genome-wide specificity. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20969–20976 (2019).

15. Bravo, J. P. K. *et al.* Structural basis for mismatch surveillance by CRISPR-Cas9. *Nature* **603**, 343–347 (2022).

16. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).

17. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).

18. Lee, J. K. *et al.* Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat. Commun.* **9**, (2018).

19. Vakulskas, C. A. *et al.* A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).

20. Schmid-Burgk, J. L. *et al.* Highly Parallel Profiling of Cas9 Variant Specificity. *Mol. Cell* 1–7 (2020). doi:10.1016/j.molcel.2020.02.023

21. Cerchione, D. *et al.* SMOOT libraries and phage-induced directed evolution of Cas9 to engineer reduced off-target activity. *PLoS One* **15**, e0231716 (2020).

22. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nature Reviews Genetics* **16**, 379–394 (2015).

23. Freeman, A. M., Mole, B. M., Silversmith, R. E. & Bourret, R. B. Action at a distance: amino acid substitutions that affect binding of the phosphorylated CheY response regulator and catalysis of dephosphorylation can be far from the CheZ phosphatase active site. *J. Bacteriol.* **193**, 4709–4718 (2011).

24. Thean, D. G. L. *et al.* Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities. *Nat. Commun. 2022 131* **13**, 1–14 (2022).

25. Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **72**, 178-186.e5 (2018).

26. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337–346 (2016).

27. Vos, P. D. *et al.* Computationally designed hyperactive Cas9 enzymes. *Nat. Commun. 2022 131* **13**, 1–11 (2022).

28. Liu, M. Sen *et al.* Engineered CRISPR/Cas9 enzymes improve discrimination by slowing DNA cleavage to allow release of off-target DNA. *Nat. Commun.* **11**, (2020).

29. Shor, O., Rabinowitz, R., Offen, D. & Benninger, F. Computational normal mode analysis accurately replicates the activity and specificity profiles of CRISPR-Cas9 and high-fidelity variants. *Comput. Struct. Biotechnol. J.* **20**, 2013–2019 (2022).

30. Frappier, V., Chartier, M. & Najmanovich, R. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Res.* **43**, W395–W400 (2015).

31. Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* **46**, W350–W355 (2018).

32. Wilf-Yarkoni, A. *et al.* Mild Phenotype of Wolfram Syndrome Associated With a Common Pathogenic Variant Is Predicted by a Structural Model of Wolframin. *Neurol. Genet.* **7**, e578 (2021).

33. Shauer, A. *et al.* Novel RyR2 Mutation (G3118R) Is Associated With Autosomal Recessive Ventricular Fibrillation and Sudden Death: Clinical, Functional, and Computational Analysis. *J. Am. Heart Assoc.* **10**, (2021).

34. Fellner, A. *et al.* In-silico phenotype prediction by normal mode variant analysis in TUBB4A-related disease. *Sci. Rep.* **12**, (2022).

35. Helbig, I. *et al.* A Recurrent Missense Variant in AP2M1 Impairs Clathrin-Mediated Endocytosis and Causes Developmental and Epileptic Encephalopathy. *Am. J. Hum. Genet.* **104**, 1060–1072 (2019).

36. Schubert, M. *et al.* Evaluate CRISPR-Cas9 edits quickly and accurately with rhAmpSeq

targeted sequencing. 1–9 Available at: www.idtdna.com.

37. Kulcsár, P. I., Tálas, A., Ligeti, Z., Krausz, S. L. & Welker, E. SuperFi-Cas9 exhibits remarkable fidelity but severely reduced activity yet works effectively with ABE8e. *Nat. Commun.* **13**, (2022).

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

39. Anderson, E. M. *et al.* Systematic analysis of CRISPR-Cas9 mismatch tolerance reveals low levels of off-target activity. *J. Biotechnol.* **211**, 56–65 (2015).

40. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).

41. Capon, S. J. *et al.* Utilising polymorphisms to achieve allele-specific genome editing in zebrafish. *Biol. Open* **6**, 125–131 (2017).

42. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018).

43. Ihry, R. J. *et al.* p53 inhibits CRISPR–Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **24**, 939–946 (2018).

44. Enache, O. M. *et al.* Cas9 activates the p53 pathway and selects for p53-inactivating mutations. *Nat. Genet. 2020 527* **52**, 662–668 (2020).

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

46. Frangoul, H. *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *N. Engl. J. Med.* **384**, 252–260 (2021).

47. Lu, Y. *et al.* Safety and feasibility of CRISPR-edited T cells in patients with refractory non-small-cell lung cancer. *Nat. Med. 2020 265* **26**, 732–740 (2020).

48. Stadtmauer, E. A. *et al.* CRISPR-engineered T cells in patients with refractory cancer. *Science* **367**, (2020).

49. Gillmore, J. D. *et al.* CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *N. Engl. J. Med.* **385**, 493–502 (2021).

50. Nahmad, A. D. *et al.* Frequent aneuploidy in primary human T cells after CRISPR-Cas9 cleavage. *Nat. Biotechnol.* **40**, 1807–1813 (2022).

51. Tsuchida, C., Brandes, N., Bueno, R., bioRxiv, M. T.- & 2023, undefined. Mitigation of chromosome loss in clinical CRISPR-Cas9-engineered T cells. *biorxiv.org*

52. Schiroli, G. *et al.* Precise Gene Editing Preserves Hematopoietic Stem Cell Function following Transient p53-Mediated DNA Damage Response. *Cell Stem Cell* **24**, 551-565.e8 (2019).

53. Kato-Inui, T., Takahashi, G., Hsu, S. & Miyaoka, Y. Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 with improved proof-reading enhances homology-directed repair. *Nucleic Acids Res.* **46**, 4677–4688 (2018).

54. Idoko-Akoh, A., Taylor, L., Sang, H. M. & McGrew, M. J. High fidelity CRISPR/Cas9 increases precise monoallelic and biallelic editing events in primordial germ cells. *Sci. Rep.* **8**, (2018).

55. Pedrazzoli, E. *et al.* An optimized SpCas9 high-fidelity variant for direct protein delivery. *Mol. Ther.* (2023). doi:10.1016/J.YMTHE.2023.03.007

56. Teruel, N., Mailhot, O. & Najmanovich, R. J. Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants. *PLoS Comput. Biol.* **17**, (2021).

57. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

58. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

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

60. Lu, X. & Olson, W. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).

61. Lu, X. & Olson, W. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **3**, 1213–1227 (2008).

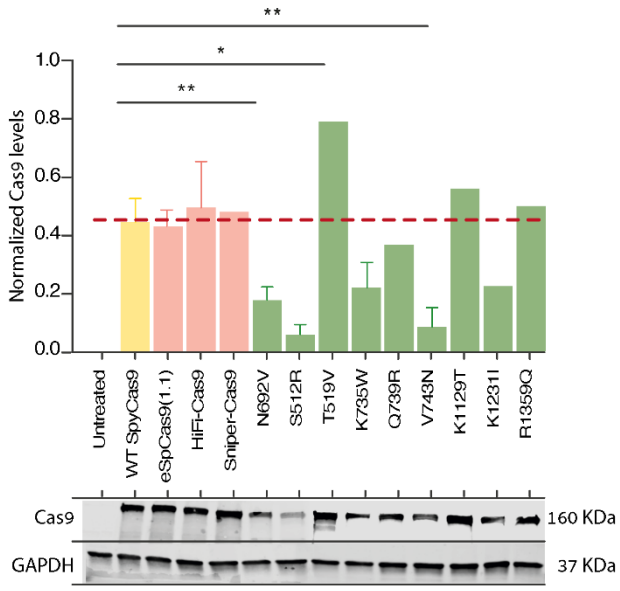
62. Zheng, G., Lu, X. & Olson, W. Web 3DNA--a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res.* **37**, (2009).

63. Colasanti, A., Lu, X. & Olson, W. Analyzing and building nucleic acid structures with 3DNA. *J. Vis. Exp.* (2013). doi:10.3791/4401

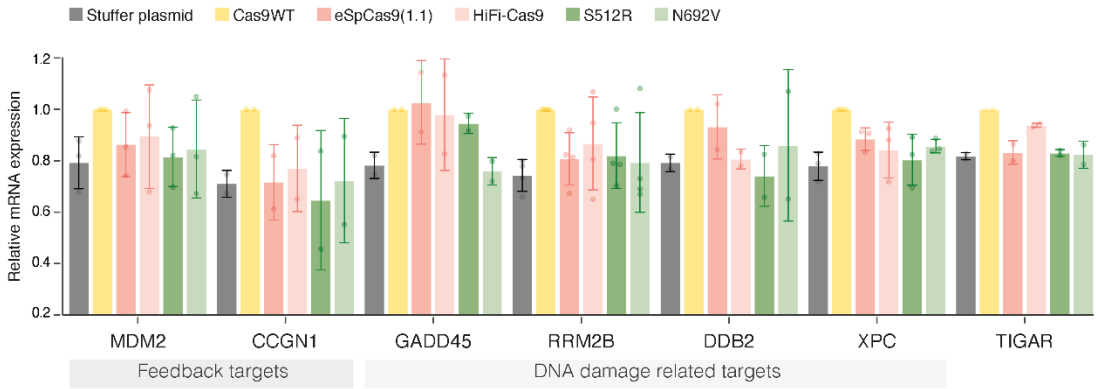
64. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).

65. Doman, J. L., Raguram, A., Newby, G. A. & Liu, D. R. Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat. Biotechnol. 2020 385* **38**, 620–628 (2020).

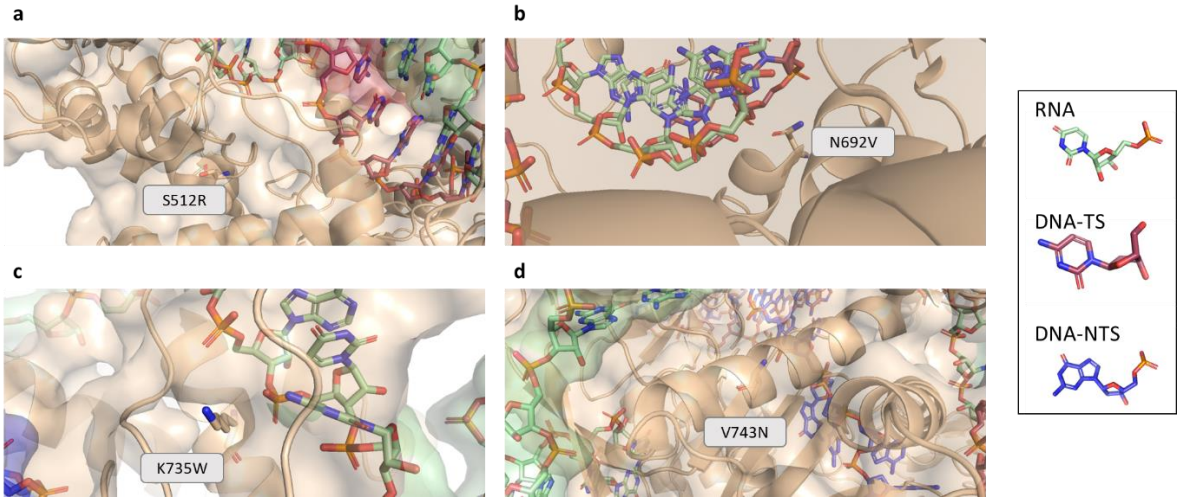
66. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).



Supplementary Figure 1: Protein expression levels of HF Cas9 variants.
Above are the protein levels of each Cas9 variant normalized to GAPDH. For WT SpyCas9 n=4, eSpCas9(1.1), HiFi-Cas9, S512R and N692V n=3, Sniper-Cas9, T519V, Q739R, K1129T, K1231I and R1359Q n=1. One-way ANOVA, Multiple comparisons to WT SpyCas9. Below are representative images of the Cas9 and GAPDH bands.



Supplementary Figure 2: Cas9 HF variants decrease p53 activation in MCF7 cells.
Introduction of Cas9 in MCF7 induces p53 activation in all tested targets. Introduction of HF variants reduces this activation.
Tested targets: MDM2 (n=3), CCGN1 (n=2), GADD45 (n=2), RRM2B (n=4), DDB2 (n=2), XPC (n=3), TIGAR (n=2).



Supplementary Figure 3: Structural description of the mutated residues of the four HF SpyCas9 variants.

RNA nucleotides are shown in green, target strand (TS) DNA in red and non-target strand (NTS) DNA in blue. The protein is represented as cartoon, while the mutated residues (**a**: S512R, **b**: N692V, **c**: K735W and **d**: V743N) are represented as sticks to visualize their distance from nucleic acid in the structure.