AI-GUIDED DATA-SCARCE ENGINEERING OF RFXCAS13D TO CREATE A CELL SELECTION TOOL

Aviv Spinner
Department of Systems Biology
Harvard Medical School

Ayush Noori
Wyss Institute for Biologically Inspired Engineering
Harvard University

Debora S. Marks
Department of Systems Biology
Harvard Medical School

George Church
Department of Genetics
Harvard Medical School

Lisa Riedmayr*
Department of Genetics
Harvard Medical School

ABSTRACT

CRISPR has revolutionized genetic engineering by providing straightforward tools for many kinds of genetic alterations. CRISPR-Cas13 is a programmable endonuclease that specifically targets and cleaves RNA. After activation by target cis-RNA binding, Cas13 also exhibits non-specific collateral trans-activity against nearby RNAs. In several conditions, this leads to apoptosis. Here, we propose to harness the collateral activity of RfxCas13d for selective induction of apoptosis in heterogeneous cell populations. We design and perform machine-guided engineering of RfxCas13d to increase collateral activity, with applications in highly specific cancer therapeutics.

1 INTRODUCTION

CRISPR and CRISPR-associated (Cas) proteins are components of a bacterial adaptive immune system that have revolutionized biology and are increasingly used for translational or therapeutic purposes like targeted genome modification and transcriptional regulation (Knott & Doudna, 2018; Anzalone et al., 2020). While most CRISPR-family enzymes cleave DNA, the class 2 type VI programmable endonuclease CRISPR-Cas13 specifically targets RNA, providing new opportunities for cell manipulation (Abudayyeh et al., 2016). Cas13 enzymes feature two Higher Eukaryotes and Prokaryotes Nucleotide-binding (HEPN) domains in a single effector molecule. After processing a CRISPR RNA (crRNA), Cas13 binds to and degrades a crRNA-complementary target RNA in a HEPN-catalyzed cleavage reaction. To date, Cas13 has been successfully used across numerous cell types and *in vivo* for highly efficient RNA knockdown, targeted RNA modification, and nucleic acid detection (Konermann et al., 2018; Kellner et al., 2019; Abudayyeh et al., 2019; Ackerman et al., 2020). After activation by target RNA binding, Cas13 also exhibits non-specific off-target collateral activity against nearby bystander RNAs. This collateral activity has been reported in both prokaryotic and eukaryotic cells, leading to an upregulation of apoptotic factors and cell death (Wang et al., 2019; Ai et al., 2022; Shi et al., 2023; Özcan et al., 2021). Thus, previous studies of Cas13 have attempted to engineer Cas13 to mitigate collateral degradation and improve Cas13 fidelity (Tong et al., 2023).

However, by increasing rather than decreasing Cas13 collateral activity, Cas13 could also be exploited for selective induction of apoptosis based on the specific transcriptomes. Efficient and precise cell subpopulation selection remains challenging, and selective induction of apoptosis based on the transcriptome holds great potential for therapeutic applications such as cancer cell elimination. Indeed, transcript expression and splicing in cancerous cells are strongly dysregulated, differing greatly from their healthy counterparts (Calabrese et al., 2020). Here, we propose to harness the collateral activity of Cas13 to develop a novel and highly specific cancer therapeutic that targets transcripts exclusively expressed in cancer cells and induces selective apoptosis. As such, we perform machine learning-guided engineering of Cas13 to increase collateral activity.

*Correspondence: lisa_riedmayr@hms.harvard.edu

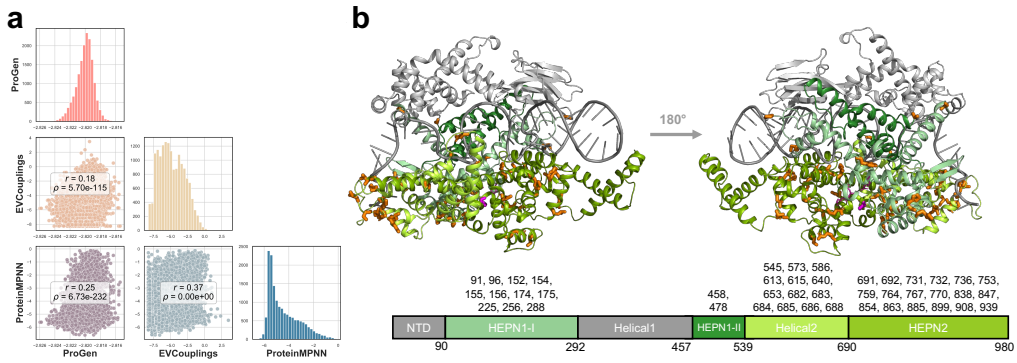

Figure 1: **Comparison of unsupervised methods and location on Cas13 structure. (a) Unsupervised model correlations of Cas13 mutations.** We learn from the whole protein universe with ProGen, local alignment with EVCouplings, and an AlphaFold-predicted structure with ProteinMPNN. Pairwise comparisons between variant effect predictions of single amino acid mutations shows that there is a small amount of correlation between predicted scores. **(b) Location of mutations on Cas13 predicted structure.** Positions selected from ProGen, EVCouplings, and ProteinMPNN models are shown in orange on the AlphaFold2 RfxCas13d predicted structure. Mutations are constrained to the HEPN and Helical2 domains.

2 Position Prioritization Approach

We first experimentally evaluated several Cas13 subtypes and crRNAs, including LwaCas13a, Pin2Cas13b, and RfxCas13d. We found that RfxCas13d, a 967-amino acid protein, was able to induce apoptosis in two cancer cell lines, HEK293T and U87-MG, as measured by Annexin V staining after 48 hours. Using prime editing in HEK293T cells, we altered two base pairs in the EGFP coding sequence. We then co-expressed RfxCas13d with a crRNA targeting the unedited EGFP transcript, thereby enhancing population-wide gene editing efficiency. This suggests that RfxCas13d selectively induces apoptosis in target transcript-expressing cells while sparing non-expressing ones. Moreover, it demonstrates the high sequence specificity of RfxCas13d, which could be exploited to distinguish single-point mutations present in cancer cells. This aligns with findings reported by Tong et al. (2023). To further increase the specific toxicity of RfxCas13d to therapeutically relevant levels, we designed a site saturation variant library on 96 selected amino acid positions of RfxCas13d. This search is guided by both the available literature and machine learning-based prioritization to select sites for mutagenesis and evaluation.

2.1 Biological Literature

We chose sites for exploration based on the literature, focusing on regions likely to influence cleavage activity. Specifically, we selected sites within the highly conserved 6 residue HEPN RXXXXH motif (Zhang et al., 2018). This motif is essential for the nuclease activity in Cas13d, and it is responsible for both the cis- and trans-RNA cleavage. Mutations in this region have been studied in other enzymes; however, no one has investigated the comprehensive set of single mutations here for RfxCas13d. In addition to these sites in the HEPN domain, we also chose to examine 40 sites that have been previously studied in Cas13x and Cas13d. Tong et al. (2023) attempted to engineer high-fidelity Cas13 variants to decrease the amount of selective toxicity. However, they also noticed several positions where selective toxicity may have been increased, and we included those positions in our assays to better understand the effects of mutations at those sites.

Conversely, we deliberately excluded regions unlikely to impact specific toxicity, such as the recognition (REC) domain. The REC domain is thought to be primarily responsible for initial recognition and binding of the target RNA. Also excluded are 16 residues from all other domains – HEPN1 and HEPN2 and Helical-2 – that are thought to be involved in crRNA binding or target RNA binding (Zhang et al., 2018).

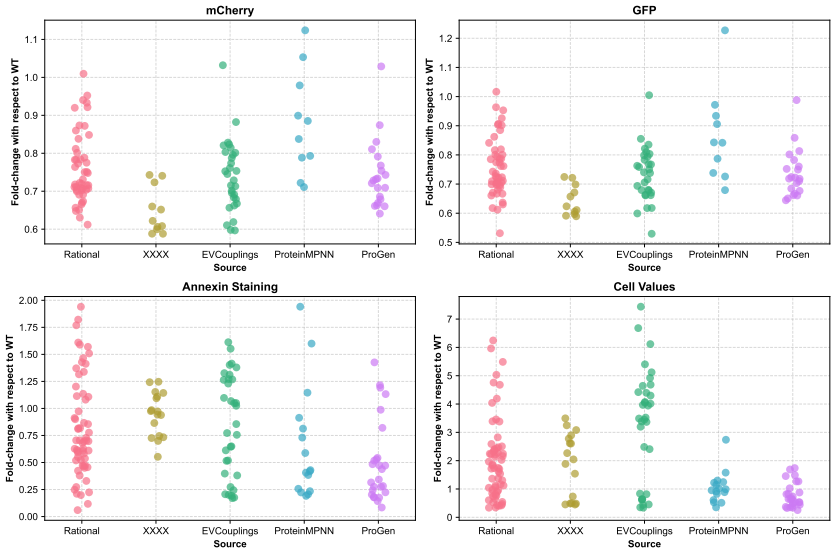

Figure 2: **Results of zero-shot engineering on four different phenotypes.** mCherry and GFP improvements are minimal and represent phenotypes closest to the engineering goal. Annexin V and cell values gains are larger but are phenotypes less relevant to the engineering goal.

2.2 Machine-Guided Prioritization

In addition to sites selected from the literature, we developed a machine learning-guided approach to expand the mutational exploration into other parts of the protein. Generally, unsupervised machine learning for proteins can learn from local sequence space (alignment-based methods), the universe of all known protein sequences (large language models), and models trained on structure. Typically, protein engineering or protein design workflows will choose one model and use that model to select which variants to test. In our case, Cas13d does not have many related sequences in the UniRef100, BFD, or MGnify databases (Suzek et al., 2007; Jumper et al., 2021; Richardson et al., 2023) and there is no resolved structure of this specific variant. In these frequent low data settings, it remains an open question of how to best leverage machine learning when no obvious choice presents itself. Here, we stratified our ML-guided prioritization across three sources of information:

(a) **Sequences:** We built a multiple sequence alignment (MSA) of 824 related Cas13 sequences from the MGnify database (Richardson et al., 2023) using 5 iterations of JackHMMER (Eddy, 2011) and used this MSA to train an EVCouplings model (Marks et al., 2011).

(b) **Structure:** AlphaFold2 (Jumper et al., 2021) was used to generate a predicted structure (1.591 Å RMSD from the closest known homolog, PDB: `6e9e` (Zhang et al., 2018)) that we used to run ProteinMPNN (Dauparas et al., 2022).

(c) **The Protein Universe:** Finally, ProGen (Madani et al., 2020), trained on the whole protein universe, was used out-of-the-box with all default parameters.

From these three models, we performed variant effect prediction on all possible single mutations. While the various model architectures and loss functions differ, we obtain the probability for any arbitrary amino acid at each site and use this as a proxy for a DMS experiment.

The predictions of each model were distinct and do not correlate well with each other (Figure 1a), suggesting that the models learn different features of the sequence landscape. Although we do not explicitly model RNA cleavage or other specific toxicity related features of Cas13, we assume that this natural property is inherently encoded within its sequence and structure. Consequently, sequences deemed more probable are expected to yield proteins with greater overall fitness. It is important to recognize, however, that computationally "improved" (*i.e.*, more probable) sequences may reflect enhancements in expression or stability rather than directly increasing specific toxic-

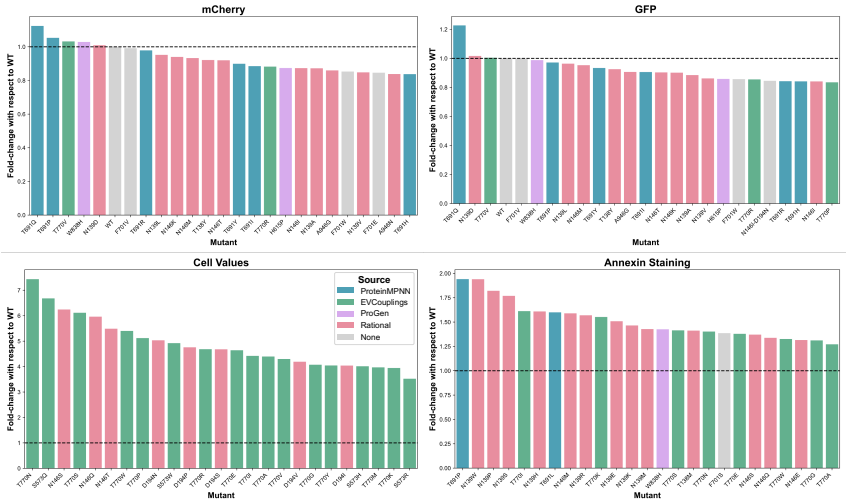

Figure 3: **Top 25 mutants per phenotype.** For competitive growth assays where mCherry and GFP are tested, `T691Q` is the top mutant and for Annexin V staining, T691P was best. Cell value changes shows the highest improvement with mutant `T770N`.

ity. However, these enhancements might indirectly influence RfxCas13d's specific toxicity and can therefore still lead to valuable improvements.

These predictions from the models were used to prioritize the positions in two ways. First, we selected the top five individual mutations from each model that were most probable. This approach is greedier and more risky since any particular mutation effect from the model has a likelihood of being incorrect. Consequently, we de-risked the positions chosen by averaging model predictions across mutations at each site, then selecting the top 14 sites with greatest average likelihood. Several of these sites overlapped between models, and the combined sites resulted in 52 ML-guided positions distributed across the protein (Figure 1b).

3 Results of Zero-Shot Variants

We synthesized a library containing all possible amino acid substitutions at the selected RfxCas13d amino acid positions identified by both literature and machine-guided prioritization, and assayed the ability of the selected variants to induce selective apoptosis in HEK293T and U87-MG cell lines. To evaluate the cell toxicity of the synthesized Cas13d mutants, a combination of cell growth and apoptosis staining was used. In this context, high AnnexinV staining indicates phosphatidylserine externalization, a hallmark of apoptosis. Cell value is another indicator of cell death: low cell value indicates that a large number of cells have been killed.

We also aimed to evaluate the selective potential of Cas13d mutants in a competetive growth assay using cell lines stably expressing either GFP or mCherry. In this assay, low GFP values indicate a strong ability of Cas13 to eliminate GFP-expressing target cells, while high mCherry levels suggest minimal toxicity against mCherry-expressing non-target cells. The mCherry levels are a proxy for measuring how many non-target cells (*i.e.*, cells not expressing the target transcript) can be enriched. This is the most relevant metric since it indicates that cells not expressing the target transcript are unharmed while the GFP-expressing target cells have been eliminated by the Cas13d enzyme. For ease of interpretation, the directionality of the metrics is held consistent such that higher is better when interpreting fold change with respect to wild-type.

The mCherry and GFP metrics from the cell growth assay were marginally improved. From the mCherry data, mutants `T691Q`, `T691P`, `T770V`, `W838H`, `and N139D` show modest enhancements over wild-type. From the GFP data that is a proxy for high toxicity in target cells, mutants `T691Q`, `N139D`, `and T770V` show improvement over wild-type. Annexin staining and cell growth values show significant improvement with a fold change improvement maximum of

	ProGen	ProteinMPNN	EVCouplings
mCherry	-0.164	0.184	0.005
Annexin Staining	-0.105	0.109	-0.086
Negative Cell Value	-0.216	0.115	0.184
Negative GFP Percent	-0.272	0.227	0.078

Table 1: **Correlation of computational and experimental predictions.** No single computational method correlates perfectly with experimental results.

$1.94\times$ and $7.44\times$ respectively. These experimental results are shown together in Figure 2. When considering the top 25 mutants of each phenotype, shown in Figure 3, mutant `T691Q` exhibits the most improvement for both mCherry and GFP whereas `T691P` is the best for Annexin staining and `T770N` is best for cell values. Generally, we observed several examples of multiple different substitutions at a single position leading to improvements, indicating loss-of-function gains. We may have observed one gain-of-function mutation, `W838H`, as defined by testing multiple different substitutions at a single position and only observing one leading to functional gains.

Notably, all of the top-performing mutants were based on ML-driven rather than literature-reported mutations: ProGen nominated position 838, ProteinMPNN nominated 691, and EVCouplings nominated 770, reflecting differences in how these models leverage evolutionary data, structural information, and statistical patterns. None of the top-performing mutations were based on previous literature, highlighting that ML-driven approaches not only recapitulate known functional sites but also uncover novel beneficial mutations that may have been overlooked in traditional studies.

Lastly, when correlating the experimentally measured values with model predictions, in Table 1 we see that none of these zero-shot models correlate particularly well with any individual phenotype. Although zero-shot models are, on their own, able to nominate several interesting mutants, they cannot predict the functions of interest.

4 Using Sparse Experimental Data for Iterative Engineering

With these few examples of labeled data, we implement several different semi-supervised strategies: Kermut (Groth et al., 2024), SaProt (Su et al., 2023), and ProteinNPT (Notin et al., 2023). These three models represent state-of-the-art semi-supervised machine learning models for protein fitness using sparse experimental data. They consistently outperformed other models on the ongoing ProteinGym leaderboard. In a similar way as described for the zero-shot predictions, we create *in silico* DMS scores for all single mutants on the wild-type background and plan to use these fitness scores to nominate mutations to test. However, unlike the previous section, we only focus experimental efforts towards the competetive growth assay mCherry and GFP phenotypes. While the Annexin V and cell value phenotypes showed more improvement and are easier to measure, the results from those assays are less useful since those phenotypes are not the engineering priority. Thus, we averaged the fold-improvement-over-WT values for mCherry and GFP as the training data, and in total we trained on 144 datapoints.

Moving forward, we will conduct 5-fold cross validation on random splits of the experimental data for these three models. If there are notable differences in how well one specific model learns the mutational landscape, we will test more single mutants from that model. Additionally, we plan on stacking the best identified mutants from either round to examine additive and synergistic effects.

5 Conclusion

This work highlights the challenges and opportunities of protein engineering when unsupervised data is limited. By leveraging both literature-guided and ML-guided approaches, we successfully identified several promising RfxCas13d variants with enhanced specific activity. Despite the zero-shot models offering insights into potential mutational hotspots, they did not consistently correlate with experimental results, reflecting the complexity of predicting functional outcomes computationally.

Using our newly-collected sparse experimental data, we plan on employing semi-supervised models such as Kermut, SaProt, and ProteinNPT to further refine RfxCas13d towards our engineering goal. These methods, which integrate limited data to guide iterative design, offer a promising framework for improving proteins with therapeutic potential. Future work will focus on stacking mutations from the initial rounds to further optimize these variants for selective cell death, with broader applications in targeted therapeutics.

REFERENCES

Omar O. Abudayyeh, Jonathan S. Gootenberg, Silvana Konermann, Julia Joung, Ian M. Slaymaker, David B. T. Cox, Sergey Shmakov, Kira S. Makarova, Ekaterina Semenova, Leonid Minakhin, Konstantin Severinov, Aviv Regev, Eric S. Lander, Eugene V. Koonin, and Feng Zhang. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, 353(6299):aaf5573, August 2016. doi: 10.1126/science.aaf5573. URL https://www-science-org.ezp-prod1.hul.harvard.edu/doi/10.1126/science.aaf5573. Publisher: American Association for the Advancement of Science.

Omar O. Abudayyeh, Jonathan S. Gootenberg, Brian Franklin, Jeremy Koob, Max J. Kellner, Alim Ladha, Julia Joung, Paul Kirchgatterer, David B. T. Cox, and Feng Zhang. A cytosine deaminase for programmable single-base RNA editing. *Science*, 365(6451):382–386, July 2019. doi: 10.1126/science.aax7063. URL https://www.science.org/doi/10.1126/science.aax7063. Publisher: American Association for the Advancement of Science.

Cheri M. Ackerman, Cameron Myhrvold, Sri Gowtham Thakku, Catherine A. Freije, Hayden C. Metsky, David K. Yang, Simon H. Ye, Chloe K. Boehm, Tinna-Sólveig F. Kosoko-Thoroddsen, Jared Kehe, Tien G. Nguyen, Amber Carter, Anthony Kulesa, John R. Barnes, Vivien G. Dugan, Deborah T. Hung, Paul C. Blainey, and Pardis C. Sabeti. Massively multiplexed nucleic acid detection with Cas13. *Nature*, 582(7811):277–282, June 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2279-8. URL https://www.nature.com/articles/s41586-020-2279-8. Number: 7811 Publisher: Nature Publishing Group.

Yuxi Ai, Dongming Liang, and Jeremy E Wilusz. CRISPR/Cas13 effectors have differing extents of off-target effects that limit their utility in eukaryotic cells. *Nucleic Acids Research*, 50(11):e65, June 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac159. URL https://doi.org/10.1093/nar/gkac159.

Andrew V. Anzalone, Luke W. Koblan, and David R. Liu. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, 38(7):824–844, July 2020. ISSN 1546-1696. doi: 10.1038/s41587-020-0561-9. URL https://www.nature.com/articles/s41587-020-0561-9. Number: 7 Publisher: Nature Publishing Group.

Claudia Calabrese, Natalie R. Davidson, Deniz Demircioğlu, Nuno A. Fonseca, Yao He, André Kahles, Kjong-Van Lehmann, Fenglin Liu, Yuichi Shiraishi, Cameron M. Soulette, Lara Urban, Liliana Greger, Siliang Li, Dongbing Liu, Marc D. Perry, Qian Xiang, Fan Zhang, Junjun Zhang, Peter Bailey, Serap Erkek, Katherine A. Hoadley, Yong Hou, Matthew R. Huska, Helena Kilpinen, Jan O. Korbel, Maximillian G. Marin, Julia Markowski, Tannistha Nandi, Qiang Pan-Hammarström, Chandra Sekhar Pedamallu, Reiner Siebert, Stefan G. Stark, Hong Su, Patrick Tan, Sebastian M. Waszak, Christina Yung, Shida Zhu, Philip Awadalla, Chad J. Creighton, Matthew Meyerson, B. F. Francis Ouellette, Kui Wu, Huanming Yang, Alvis Brazma, Angela N. Brooks, Jonathan Göke, Gunnar Rätsch, Roland F. Schwarz, Oliver Stegle, and Zemin Zhang. Genomic basis for RNA alterations in cancer. *Nature*, 578(7793):129–136, February 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-1970-0. URL https://www.nature.com/articles/s41586-020-1970-0. Number: 7793 Publisher: Nature Publishing Group.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/

10.1126/science.add2187. Publisher: American Association for the Advancement of Science.

Sean R. Eddy. Accelerated Profile HMM Searches. *PLOS Computational Biology*, 7(10):e1002195, October 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002195. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195. Publisher: Public Library of Science.

Peter Mørch Groth, Mads Herbert Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. November 2024. URL https://openreview.net/forum?id=jM9atrvUii.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Number: 7873 Publisher: Nature Publishing Group.

Max J. Kellner, Jeremy G. Koob, Jonathan S. Gootenberg, Omar O. Abudayyeh, and Feng Zhang. SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nature Protocols*, 14(10):2986–3012, October 2019. ISSN 1750-2799. doi: 10.1038/s41596-019-0210-2. URL https://www.nature.com/articles/s41596-019-0210-2. Number: 10 Publisher: Nature Publishing Group.

Gavin J. Knott and Jennifer A. Doudna. CRISPR-Cas guides the future of genetic engineering. *Science*, 361(6405):866–869, August 2018. doi: 10.1126/science.aat5011. URL https://www.science.org/doi/abs/10.1126/science.aat5011. Publisher: American Association for the Advancement of Science.

Silvana Konermann, Peter Lotfy, Nicholas J. Brideau, Jennifer Oki, Maxim N. Shokhirev, and Patrick D. Hsu. Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell*, 173(3):665–676.e14, April 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.02.033. URL https://www.cell.com/cell/abstract/S0092-8674(18)30207-1. Publisher: Elsevier.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. ProGen: Language Modeling for Protein Generation. preprint, Synthetic Biology, March 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.03.07.982272.

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE*, 6(12):e28766, December 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0028766. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766. Publisher: Public Library of Science.

Pascal Notin, Ruben Weitzman, Debora S. Marks, and Yarin Gal. ProteinNPT: Improving Protein Property Prediction and Design with Non-Parametric Transformers, December 2023. URL https://www.biorxiv.org/content/10.1101/2023.12.06.570473v1. Pages: 2023.12.06.570473 Section: New Results.

Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A Gurbich, Varsha Kale, Anton Korobeynikov, Shriya Raj, Alexander B Rogers, Ekaterina Sakharova, Santiago Sanchez, Darren J Wilkinson, and Robert D Finn. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1080. URL https://doi.org/10.1093/nar/gkac1080.

Peiguo Shi, Michael R. Murphy, Alexis O. Aparicio, Jordan S. Kesner, Zhou Fang, Ziheng Chen, Aditi Trehan, Yang Guo, and Xuebing Wu. Collateral activity of the CRISPR/RfxCas13d system in human cells. *Communications Biology*, 6(1):1–8, March 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04708-2. URL https://www.nature.com/articles/s42003-023-04708-2. Number: 1 Publisher: Nature Publishing Group.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary. October 2023. URL https://openreview.net/forum?id=6MRm3G4NiU.

Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm098. URL https://doi.org/10.1093/bioinformatics/btm098.

Huawei Tong, Jia Huang, Qingquan Xiao, Bingbing He, Xue Dong, Yuanhua Liu, Xiali Yang, Dingyi Han, Zikang Wang, Xuchen Wang, Wenqin Ying, Runze Zhang, Yu Wei, Chunlong Xu, Yingsi Zhou, Yanfei Li, Minqing Cai, Qifang Wang, Mingxing Xue, Guoling Li, Kailun Fang, Hainan Zhang, and Hui Yang. High-fidelity Cas13 variants for targeted RNA degradation with minimal collateral effects. *Nature Biotechnology*, 41(1):108–119, January 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01419-7. URL https://www.nature.com/articles/s41587-022-01419-7. Number: 1 Publisher: Nature Publishing Group.

Qixue Wang, Xing Liu, Junhu Zhou, Chao Yang, Guangxiu Wang, Yanli Tan, Ye Wu, Sijing Zhang, Kaikai Yi, and Chunsheng Kang. The CRISPR-Cas13a Gene-Editing System Induces Collateral Cleavage of RNA in Glioma Cells. *Advanced science (Weinheim, Baden-Wurttemberg, Germany)*, 6(20):1901299, October 2019. ISSN 2198-3844. doi: 10.1002/advs.201901299. URL https://europepmc.org/articles/PMC6794629.

Cheng Zhang, Silvana Konermann, Nicholas J. Brideau, Peter Lotfy, Xuebing Wu, Scott J. Novick, Timothy Strutzenberg, Patrick R. Griffin, Patrick D. Hsu, and Dmitry Lyumkis. Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. *Cell*, 175(1):212–223.e17, September 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418311735.

Ahsen Özcan, Rohan Krajeski, Eleonora Ioannidi, Brennan Lee, Apolonia Gardner, Kira S. Makarova, Eugene V. Koonin, Omar O. Abudayyeh, and Jonathan S. Gootenberg. Programmable RNA targeting with the single-protein CRISPR effector Cas7-11. *Nature*, 597(7878):720–725, September 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03886-5. URL https://www.nature.com/articles/s41586-021-03886-5. Publisher: Nature Publishing Group.