Proposal for PhD Research: Quantifying the Uncertainty in both On and Off-target Activity of Prime Editing

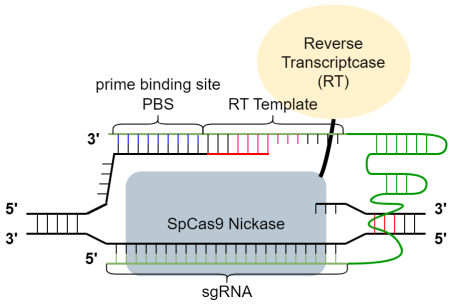
Prime editing is a novel gene editing technology that enables theoretically any base substitution, insertion, and deletion without the need for double-strand breaks or donor DNA templates[1, 2]. Shown in Figure 1a, the prime editors's versatility can mostly be attributed to the reverse transcription mechanism. After the Cas9 nickase cleaves the target DNA strand, the PBS (prime binding site) portion of the pegRNA (prime editing guide RNA) hybridizes with the now floating 3' end of the target DNA strand. The reverse transcriptase (RT) then extends the 3' end of the target DNA strand using the RT template sequence in the pegRNA. The resulting DNA-RNA hybrid is then resolved by the endogenous DNA repair machinery, resulting in the desired edit.

Due to its superior capability, prime editors have the potential to cure up to 90% of diseases resulted from genetic mutations. However, its clinical application is still limited by the frequently low on target efficiency and the potential off-target activity[3].

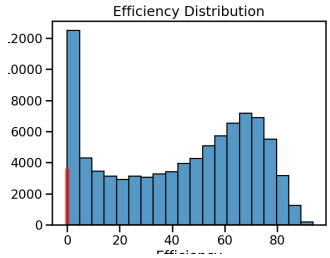
My master's thesis focused on predicting the on-target activity of prime editing using ensemble learning techniques and the Transformer model to improve pegRNA guide design qualities and thus overall efficiency, and achieved state-of-the-art performance on several existing datasets. As a result, moving on to the prediction of off-target activity of prime editing is a natural next step.

With the recent development of a number of off-target site detection protocols supporting prime editors, it is now more and more possible to quantify their off target activity using big data and deep learning techniques[4, 5]. A number of off-target prediction tools have been developed for CRISPR-Cas9, which makes up part of the prime editing system, while DeepPrime-Off developed via fine tuning DeepPrime on the off target sites remains as one of the few prime editing off-target prediction tools. Although the CRISPR-Cas9 system was an essential part of the prime editing system, as studies have pointed out, the reverse transcription process involving the PBS and RT sequences can also impact the off-target activity. Thus, novel tools are needed to produce more accurate predictions the off-target activity of prime editing[4, 6].

On top of extending the model to predict the off-target activity of prime editing, I propose to modify the models to output the probability distribution of the editing efficiency, rather than a point estimation.



(a) Prime editing process



(b) Distribution of measured editing efficiency

Although majority of the existing solution (including my master's thesis) was aiming at producing a point estimation as the predicted editing efficiency[7, 6, 8], the prime editing process itself can be viewed as a stochastic process due to the complexity of the physical processes involved. Thus, it could be more informative to model the posterior distribution of the activity of prime editing given the target loci and pegRNA sequence. For examples, researchers may be able to select the pegRNA sequence that has lower variance in the predicted editing efficiency, if multiple guides showed similar mean editing efficiency.

Recently, a probabilistic deep learning model, crispAI, was proposed to fit a zero inflated negative binomial distribution to the off-target activity of CRISPR-Cas9 on each potential target site[9]. It was able to accuracy predict the off-target activity in both minimal activity and high activity sites ....

To validate the probabilistic approach, I conducted some preliminary experiments on the on-target prediction of prime editing efficiency using the mathis et al's 90k PE2 HEK293T dataset[10].

(Add details of the experiment and results)

Additional models including MC dropout/Deep ensemble

Overall, with the on and off target activity of prime editing quantified, we can provide a complete overview of the outcome of using a specific pegRNA sequence on a specific target loci in a specific cell line. This should noticeably improve the safety and efficiency of prime editing, and thus accelerate its clinical application.

References

[1] Liu David R. et al. "Search-and-Replace Genome Editing without Double-Strand Breaks or Donor DNA". In: *Nature* 576.7785 (Dec. 5, 2019), pp. 149–157. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1711-4. URL: https://www.nature.com/articles/s41586-019-1711-4 (visited on 02/08/2024).

[2] David R. Liu and Peter J. Chen. "Prime Editing for Precise and Highly Versatile Genome Manipulation". In: *Nature Reviews Genetics* 24.3 (Mar. 2023), pp. 161–177. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-022-00541-1.

URL: https://www.nature.com/articles/s41576-022-00541-1 (visited on 02/16/2024).

[3] Zhihan Zhao et al. "Prime Editing: Advances and Therapeutic Applications". In: *Trends in Biotechnology* 41.8 (Aug. 1, 2023), pp. 1000–1012. ISSN: 0167-7799, 1879-3096. DOI: 10.1016/j.tibtech.2023.03.004. pmid: 37002157. URL: https://www.cell.com/trends/biotechnology/abstract/S0167-7799(23)00085-9 (visited on 10/16/2024).

[4] Shun-Qing Liang et al. "Genome-Wide Profiling of Prime Editor off-Target Sites in Vitro and in Vivo Using PE-tag". In: *Nature Methods* 20.6 (June 2023), pp. 898–907. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01859-2. URL: https://www.nature.com/articles/s41592-023-01859-2 (visited on 09/30/2024).

[5] Ming Zhu et al. "Tracking-Seq Reveals the Heterogeneity of off-Target Effects in CRISPR–Cas9-mediated Genome Editing". In: *Nature Biotechnology* (July 2, 2024), pp. 1–12. ISSN: 1546-1696. DOI: 10.1038/s41587-024-02307-y. URL: https://www.nature.com/articles/s41587-024-02307-y (visited on 09/28/2024).

[6] Goosang Yu et al. "Prediction of Efficiencies for Diverse Prime Editing Systems in Multiple Cell Types". In: *Cell* 186.10 (May 2023), 2256–2272.e23. ISSN: 00928674. DOI: 10.1016/j.cell.2023.03.034. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867423003318 (visited on 05/03/2024).

[7] Nicolas Mathis et al. "Machine Learning Prediction of Prime Editing Efficiency across Diverse Chromatin Contexts". In: *Nature Biotechnology* (June 21, 2024), pp. 1–8. ISSN: 1546-1696. DOI: 10.1038/s41587-024-02268-2. URL: https://www.nature.com/articles/s41587-024-02268-2 (visited on 06/23/2024).

[8] Jonas Koeppel et al. "Prediction of Prime Editing Insertion Efficiencies Using Sequence Features and DNA Repair Determinants". In: *Nature Biotechnology* 41.10 (Oct. 2023), pp. 1446–1456. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-023-01678-y. URL: https://www.nature.com/articles/s41587-023-01678-y (visited on 02/07/2024).

[9] Furkan Özden and Peter Minary. "Learning to Quantify Uncertainty in Off-Target Activity for CRISPR Guide RNAs". In: *Nucleic Acids Research* (Sept. 14, 2024), gkae759. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkae759. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkae759/7757940 (visited on 09/28/2024).

[10] Nicolas Mathis et al. "Predicting Prime Editing Efficiency and Product Purity by Deep Learning". In: *Nature Biotechnology* 41.8 (Aug. 2023), pp. 1151–1159. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-022-01613-7. URL: https://www.nature.com/articles/s41587-022-01613-7 (visited on 04/24/2024).