nature portfolio

Peer Review File

Custom CRISPR-Cas9 PAM variants via scalable engineering and machine learning

Corresponding Author: Dr Benjamin Kleinstiver

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
In this manuscript, Silverstein et al. described a framework of engineering PAM-altered SpCas9 proteins with help of machine learning (ML). By shortlisting functional candidates from a library of PI-domain-engineered SpCas9 variants and characterising their PAM preferences, the full mutational space was explored through ML, resulting in a model termed as PAMmla. While the ML method used here does not sound new, PAMmla identified various unique clusters of Cas variants with altered PAM selections that were more specific and efficient than the original training data. These novel variants were capable of editing various genomic loci with relatively lower off-target events while maintaining similar or higher activity level than reported PAM-relaxed Cas variants. By an in silico directed evolution (ISDE) approach, the full list of 64M SpCas9 variants included in PAMmla can be navigated with user-defined parameters with reduced computing power while still enabling selection of decently performed variants. By demonstrating the possibilities of predicting PAM-altered yet restricted Cas9 variants by PAMmla, the authors provided new insights in tailored Cas design for applications, in which PAM-relaxed variants were not usually favourable.

Here are some comments and questions for the authors:

1. The authors suggested that they have “obtained 655 unique SpCas9 enzymes that survived the selections (Sup. Table 1).” (line 135). While there are 655 rows of entries labeled as “bacterial selections”, only 634 of them were unique AA entries. For example, the enzyme “MQKSER” appeared 4 times, all labeled as “bacterial selections on NGCA PAM”, but with slightly different values in the measurements. The authors should confirm the accuracy of the reported numbers and the table.

2. In the same table (Sup. Table 1), there are quite a lot of identical values in 14 decimal places (e.g. -8.46813474563119 appeared for 2752 times; -7.4827558925092 appeared for 3807 times etc.), are they some sort of defects or they are indeed the correct values to appear in those boxes? Overall, it is not very clear in how to interpret the table values.

3. While transformation efficiency and coverage of the 6AA_NNS library was not stated, “~24-48 colonies” (line 697) were selected by the authors from each selection for downstream characterisation using HT-PAMDA. It was however uncertain how representative 24-48 colonies can be in such context. This was essential to evaluate if the claims “an enzyme’s most efficiently targeted PAM did not always correlate with the PAM on which that variant was selected (Fig. 1e, Sup. Figs. 4 and 5).” (line 203-204); “PAMmla resulting variants were more specific for the target PAM than those obtained from bacterial selections” (line 313-315) were resulted from sampling biases or the intrinsic issue of bacterial selection.

4. The authors mentioned that they compared different ML models, but the description of the ML methods used are very limited. Is this based on the published MLDE method, or any novel models were developed?

5. Following from the authors’ comments on the variants recovered from positive selection (line 311-313 and Fig 3f), the authors may want to indicate (in terms of activities ranking) where these positively selected variants are predicted to be among the full library (64 millions)? Are they at the top 10% range of the library? Should the authors expect to find more of the maximum efficiency enzymes predicted by PAMmla if they increase the stringency of the selection?

6. Given the HT-PAMDA results of the selected variants from the bacteria screen, it was likely that variants could appear in multiple PAM groups and should naturally enriched more in PAMs that they edited more efficiently during bacteria selection.

Labeling where these variants were selected according to the enrichment scores compared to the control plate would be a more sensible and fair way to depict an inferior position of bacteria screening approach.

7. Only one of the spacer sequences used in HT-PAMDA was the same as the one used in bacteria selection. Would that set of data show more consistent result with the bacteria selection in terms of the most efficient PAMs? The authors should address the potential spacer discrepancy as well as the different environmental factors during bacteria screen and in vitro assay.

8. There are discrepancies on PAM specificity based on the bacterial selection versus the in vitro assay HT-PAMDA. It is not certain whether such discrepancies also exist when comparing HT-PAMDA and the assay results obtained from the experiments in mammalian cells. The authors should establish that correlation to confirm such consistency (similar to Fig. 3c, but with mammalian cell validation data). In mammalian cells, are the PAM-altered variants tested (for example those in Fig. 4b) not cutting the other PAMs (among the 16 NGNN PAMs) that they were predicted not to?

9. Currently the off-targeting analysis of the MRRWMR enzyme was done in two separate sgRNAs instead of the one used in the RHO P23H locus, which it was designed to target. It would be helpful to also perform GUIDE-seq2 on this sgRNA to properly match the reason why this variant was selected in the first place.

10. In Figure 6g, while “reads containing indels that span the P23H mutation, edited counts were distributed using the ratio of WT to mutant as observed for the identifiable edited reads”, the authors should also indicate how much of the edits contained such indels for better evaluation of the data. The assumption to distribute them in ratio of WT to mutant may not necessarily correct but the effect would be negligible if such event was rare to start with.

11. It would be great if the reason why the KRHWMR enzyme can specifically edit NGTG but not NGGG as demonstrated in Figure 6g can be speculated through structure modeling. This could further strengthen the idea of allele-specific editor design using the PAMmla and provide more insight on PAM-altered Cas design.

12. The authors suggested the use of ISDE over direct sorting approach when handling the 64M prediction list due to its lower computing demand yet still being able to select desired variants. The authors should benchmark the two methods in these two aspects to demonstrate that ISDE was indeed faster to work with and capable of selecting the variants comparable to the exhaustive sorting approach. This should further demonstrate that ISDE was the more favourable option.

13. Figure 6 c-f indicated that the ISDE regime would identify enzymes with quite different maximum activities in each trajectory. In general, how many ISDE runs the authors would recommend to ensure that one may find the optimal variant for experimental application? Will there be different rounds of ISDE iteration based on the position of the SNP on the PAM?

14. Supplementary Note 5 claims that active variants typically have PAM activity correlations of 0.7 to 0.9. The authors should elaborate on that and plot such correlations among PAM sequences. The authors should also elaborate about the definition of "inactive enzymes" here, i.e. how many inactive PAM sequences (cut off k > 10^-4?) an inactive enzyme has to have in order to achieve such correlation of 0.1?

Minor questions:
15. Fig 3c and S9b, h, and g: Are the predicted and empirical k values derived from 1 specific PAM sequence or averaged across many PAM sequences?

16. Figure S9d and f: Can the authors explain on how the breakdown values were calculated?

17. The detailed GUIDE-seq2 data were currently not available for most of the GUIDE-seq2 experiments. Please provide them either as supplementary tables or visualised figures as shown in Supp. Figure 19b.

18. Supplementary Note 5: Where are Sup. Fig. 7g and 7h?

19. Single guide RNA (sgRNA) might be a more proper term than just guide RNA (gRNA) as the latter was used to describe the crRNA + tracrRNA complex, which was not linked by a linker as the modern sgRNA that was used in this study.

(Remarks on code availability)
There is no supporting github codes provided in the article. The ML seems to based on the published MLDE framework (Wittmann et al., Cell Syst., 2021) (i.e., it is not very clear as only very limited description on the ML method used is provided). If so, such MLDE framework for CRISPR engineering was previously demonstrated (Thean et al., Nature Comm, 2022), which is also cited by the authors.

Referee #2

(Remarks to the Author)
SpCas9 is the most popular gene editing tool and numerous approaches have been utilized to engineer this protein for optimal target selectivity and specificity. One key factor is its requirement for the specific detection of a PAM sequence which limits the range of available target sites. Previous research focused on the engineering of the PAM interacting (PI) domain, resulting in commonly used Cas9 variants with relaxed PAM recognition or altered PAM requirements. In the present

manuscript, Silverstein et al. built on this concept, assayed a saturation mutagenesis library and used this data to train a neural network to identify novel Cas9 variants with unique PAM preferences. Successful examples of this approach are described as selected Cas9 variants were identified to exhibit novel PAM preferences.

In general, this is a highly relevant topic with a lot of competing research activities. Machine learning was for example used to optimize sgRNA selection (10.1016/j.celrep.2024.113765), but also applied for the detection of novel Cas9 PAM-interacting domains (doi.org/10.1371/journal.pcbi.1011621). Alternative approaches to acquire a larger selection of specific PAM requirements include the screening of natural Cas9 variants from other host bacteria or the generation of Cas9 variants without PAM requirements. The authors provide convincing arguments against the latter approach.

The focus of this manuscript are six amino acids in the PI-domain of SpCas9 that were subjected to saturation mutagenesis. These residues were specifically selected as their alterations was previously shown to alter PAM specificity. Here, the authors characterized the PAM requirements of SpCas9 enzymes with variant PI domains using HT-PAMDA. The investigated PAM is (N)GNN and I initially wondered why the G at the second position was fixed for the training set as (N)NN would also result in 16 screenable experiments. This is detailed in supplementary note 3, but I would find it important to indicate in the main text that the second position could not be changed due to its interactions with a network of amino acids coordinated by R1333. In theory, being able to change the second PAM position would be more desirable than changing only the third and a "new" fourth position. The rationale for this PAM library should also be shortly discussed outside of the supplementary note.

The bacterial-based selection approach provided 655 Cas9 variants with unique PI domains and revealed that most enzymes utilized a canonical NGG PAM or a relaxed PAM recognition, indicating that the PI domain co-evolved with PAM elements using regions outside of the selected six amino acids. I agree with authors' discussion point that it would be necessary to expand the six amino acid library in the future.

As 18% of 135 randomly mutated versions (without selection) provided (non-canonical) PAM preferences, the authors then selected a machine learning approach termed PAMmla to predict quantitative rate constants (ks) for Cas9 variations for each of the 64 (N9)GNN PAM versions. I am not an expert on machine learning and will refrain from reviewing this portion of the manuscript in detail. However, I would be curious how one measures that the training input was of sufficient quality. Most enzyme variants (82%, according to the 135 randomized versions) should result in versions that are not capable of editing any of the investigated PAM sequences. Should an ideal training set then not include 82% of inactive variants to better evaluate amino acid combinations that result in non-functional enzymes? What happened if this correction approach was not applied: "To account for the imbalanced nature of our training data towards NGGN and NGTN enzymes and those preferring a G at the 4th 237 PAM position (Fig. 1f), we assigned a label to each training example based on its most active PAM and randomly over-sampled to balance across PAM classes"?

The authors determined PAM profiles for 253 novel Cas9 versions and verified editing efficiencies for PAM elements that have not been described before. In general, the performance of PAMmla is impressive and predicted PAM requirement could be confirmed. Interestingly, the generated enzymes usually showed specific PAM requirements and not only relaxed PAM recognition. I am wondering if structure modelling is able to explain this feature. Here, it would be desirable to analyse the modelled structure of PI domain variants (...for each of the HT-PAMDA profile clusters) and dock DNA targets with specific PAM versions. This approach should help to validate the authors' results and also stimulate further structure-based engineering of amino acids near the six mutated residues.
The manuscript is written in clear and coherent fashion and I only disliked that crucial information is "hidden" in eight supplementary notes. A reference to this paper (Malbranke, C et al. (2023) PLoS Comput Biol 19(11): e1011621) would be appropriate to indicate an alternative ML approach to generate Cas9 variants with modified PAM specificity.

Few references need to be updated to be able to follow the methods section and the data availability section will need to be adjusted when primary data is available:
- Line 821: As described elsewhere (Hille LT, Christie KA, Walton RT, et al., submitted)
- line 867: GUIDE-seq2 reactions were performed essentially as described (Lazzarotto & Li et al, in preparation) with minor modifications)
- line 889: a summary of GUIDE-seq2 data will be made available in Sup. Table 6
- several instances: (to be deposited on Github).

(Remarks on code availability)

Referee #3

(Remarks to the Author)
In this work the authors gather a unique dataset on the activity of Cas9 variants on different PAM sequences. A mutant library with randomized amino acids at 6 important positions was generated and first screened using a bacterial selection system on each of the 16 possible NGNN PAMs. 655 Cas9 variants were then cloned and their PAM specificity profiled using the HT-PAMDA method. Interesting variants were selected which provided insights into PAM binding specificity determinants, but the process was not satisfactory in terms of isolating efficient and specific variants against non-canonical PAMs.

The authors then generated additional data on 135 randomly chosen variants and used the total dataset to train models able predict PAM specificity profiles based on the 6 residues modified in the library. Their PAMmla model achieves successful predictions of PAM profiles and was then used in the design of variants able to target specific loci efficiently. 253 active variants were generated and their activity measured. The authors further successfully test PAMmla predicted nucleases and base editors in human cells, showing activity higher than the relaxed SpG-Cas9 protein as well as more specificity. Finally the authors generate variants with the goal to discriminate between two closely related PAM sites. Here too the model is successful at identifying a variant specific for the desired PAM.

Altogether the authors generated a very impressive amount of high-quality data which provides important novel insights into the determinants of Cas9 PAM specificity. They convincingly show that this novel dataset can be used to train models that are useful in the design of novel Cas9 variants. While I applaud the authors for the quality of the experimental work, which will be an important ressource for the community, I have a few important criticisms.

The authors demonstrate improved activity versus the PAM-relaxed enzyme SpG, but not against natural Cas9 variants. Is there still a margin of improvement here in term of activity and specificity? How does the engineering of SpCas9 variants compare to the use of natural Cas9 variants with different PAM specificities or the construction of chimeric Cas9 variants that swap the PI domain entirely with that of natural variants?

Importantly, I find the modeling aspects of the manuscript lacking. This is concerning given that the angle taken by the authors in their writing is putting the machine learning part forefront.

The modeling methods are not described in sufficient details and could be performed to a higher standard. How were the model hyperparameters chosen? Did the authors conduct an hyperparameter search? The network used is already fairly small, but could an even smaller model work as well, in particular considering that the size of the training set is fairly small? On that note, what was the total size of the dataset used? Is it 655 + 135 ? It should be much easier to find this information in the manuscript.
The authors perform some comparison between their neural network model, a linear model and a random forest model, however it seems like no effort was made to encode interaction features in the linear model. Encoding all pairwise interactions and fitting a linear model with regularization is likely to provide similar performance to the neural network model while being a lot easier to interpret. Such feature encoding could also benefit decision tree based models, for which methods such as SHAP (SHapley Additive exPlanations) can provide interpretable insights.
Could the authors also perform some analysis of the features captured by the model to predict active Cas9 enzymes?

I don't understand why the ISDE process is necessary? The authors have already computed the PAMmla predictions for all the 64 million sequences. Computing a score on these should be very fast and sorting 64 million numbers should take at most a few hours and possible much less using fast / parallel sorting algorithms, which would give a perfect knowledge of the landscape rather than requiring an optimization algorithm.

Structural modeling of the variants obtained (at least the most interesting ones) would strengthen the manuscript.

Finally, one of the main take away message from the manuscript is that generating high quality data on mutant libraries can help model fitness landscapes and design improved protein variants. While the impact of the work is made higher because it was done on Cas9, generating possibly useful variants, this message in itself is not novel and similar principles have been applied to numerous proteins, including model proteins such as GFP (https://www.nature.com/articles/s41467-023-38099-z) and antibodies (e.g. https://www.nature.com/articles/s41467-023-39022-2 to take just one example).

Minor comments:
The authors should discuss and cite related work including: https://doi.org/10.1371/journal.pcbi.1011621
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9039034/
https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.full.pdf

Initial screening in the bacterial work is restricted to NGNN PAMs. This choice should be better explained.
The authors should discuss and cite related work by Malbranke et al., PLoS Comp. Biol. 2023 (https://doi.org/10.1371/journal.pcbi.1011621) as well as preprinted work by Ruffolo et al. (https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.full.pdf) if the journal allows.

(Remarks on code availability)

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
In this revised manuscript, Silverstein et al. have addressed and strengthened the proof for the key questions, including the engineering framework using ML and ISDE methods, and the structural analysis of the SpCas9 enzyme. The modified

version is significantly improved and of high quality. There is one additional question here that I would like to raise.

1. The number of enzymes analyzed for PAMmla-predicted enzymes was insufficient to demonstrate that these enzymes with altered PAMs can reduce genome-wide off-target effects compared to PAM-relaxed enzymes. Current results were based on five enzymes, which only show that some PAMmla enzymes with altered PAMs can reduce off-target effects. To fully assess the application of PAMmla on the specificity of the SpCas9 enzyme, a larger number of enzymes (at least 10-20, ideally over 50-100) would need to be analyzed. The authors may want to revise the related statements/claims to ensure validity.

(Remarks on code availability)
The PAM activity prediction website (https://pammla.streamlit.app) was working but is sleeping due to inactivity

Referee #2

(Remarks to the Author)
The authors submitted a revised version of their manuscript that adequately addressed my previous comments. They discuss effects of varying the 2nd PAM position and added an analysis to investigate the importance of oversampling. I enjoyed the added sections on structure modelling approaches to understand the molecular basis for altered PAM requirement. Finally, two references for very recent papers with alternative ML approaches to generate Cas9 variants with modified PAM specificity have been added. The citation of the unpublished GUIDE-seq2 work and the final data availability statement will need to be discussed with the Nature editorial board.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The authors have satisfactorily addressed my all comments. The manuscript is now much better than the initial submission. I support publication and congratulate the authors on a very nice piece of work. My only remaining concern might be with the title which in my opinion places too much emphasis on machine learning, which is not particularly original in this study, and not enough on the dataset which is the key contribution.

(Remarks on code availability)
I didn't try to run the code myself, but it seems to be well documented and a useful ressource for the community.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have sufficiently addressed my comments and I have no further questions for the authors. Congratulations on the great piece of work!

(Remarks on code availability)

Introduction to resubmission:

We are pleased to resubmit our manuscript that describes the development and application of PAMmla. We thank the Referees for their time and effort reviewing our initial submission, and for their critiques and suggestions to perform experiments that have collectively strengthened our study. Below we provide our responses to the Referees' comments in **blue text**, and for convenience have included some of the new or modified figures in our response. Our resubmission includes revisions throughout the manuscript and supplementary materials that are similarly **highlighted with blue text.**

We appreciate the Referees' kind feedback and positive comments about our work describing it as: "*a highly relevant topic*", that "[the authors] *identified various unique clusters of Cas variants with altered PAM selections that were more specific and efficient than the original training data*", "*They convincingly show that this novel dataset can be used to train models that are useful in the design of novel Cas9 variants*", and "*the authors generated a very impressive amount of high-quality data which provides important novel insights into the determinants of Cas9 PAM specificity*". The Referees also noted that "*the performance of PAMmla is impressive*", the "*manuscript is written in clear and coherent fashion*", and that the "*PAMmla model achieves successful predictions of PAM profiles*".

Based on the Referees' constructive feedback, we have revised our manuscript to include several new datasets that we hope address their main concerns. Some of the major modifications to our manuscript include:

- Improved descriptions of the ML methods used & exploration of additional features: We have expanded our description of the ML models that we utilized in the Methods section and have also provided the source code for the PAMmla model on GitHub. Based on Referee feedback we also trained a linear model using pairwise amino acid combinations, which improved performance nearly to the level of the neural network, demonstrating that alternative models can effectively be used to predict the relationship between SpCas9 amino acid sequence and PAM preference (**Sup. Fig. 9a**).
- Structural modeling to investigate novel preferences of PAMmla predicted enzymes: We performed structural modeling of common mutations found in PAMmla predicted enzymes, forming new hypotheses about how novel AA:PAM contacts mutated new PAM specificities. These results are extensively discussed in our manuscript, in **Figs. 6h,i**, **Sup. Figs. 15a-l** and **30a,** and **Sup. Notes 6** and **10**.
- Analysis of ML feature importances contributing to PAM predictions: We used SHapely Additive exPlanations (SHAP) to interpret the features leading to PAMmla predictions. This provided insights into the importance of SpCas9 amino acid substitutions that contribute to novel PAM specificities (**Sup. Fig. 14** and **Sup. Note 6**), and into the mechanism by which allele-specific enzymes MRRWMR and KRHWMR distinguish their preferred and non-preferred PAMs (in **Figs. 6j,k**, **Sup. Figs. 30b-e,** and **Sup. Note 10)**.
- Guidelines & benchmarking of the ISDE method: Assessment of PAMmla-ISDE to identify optimal enzymes revealed that ISDE can recover the most active enzyme without the need for full sorting or downloading the full predictions dataset (**Sup. Figs. 23b-d**), demonstrating the utility of ISDE to explore PAMmla predictions. We additionally explored how varying ISDE parameters can impact the fraction of top enzymes recovered and provide guidelines for ISDE parameters.
- Additional characterization of therapeutically relevant enzymes: We further characterized the PAMmla-predicted allele-specific enzymes MRRWMR and KRHWMR for targeting the *RHO* P23H mutation, including *in vivo* demonstration that these enzymes are effective and allele-specific in heterozygous humanized *RHO* P23H mice (**Fig. 6h** and **Sup. Figs. 29a,b**). Off-target analysis via GUIDE-seq2 revealed that MRRWMR and KRHWMR both substantially improved editing safety by reducing genome wide off-targets compared to SpG and SpRY (**Fig. 6g** and **Sup. Figs. 28c-f**).
- Comparison of PAMmla enzymes to Cas9 orthologs: We compared naturally occurring Cas9 orthologs predicted to distinguish the *RHO* P23H alleles with our PAMmla predicted enzymes. Testing of three Cas9 orthologs failed to generate detectable indels at the *RHO* P23H target site (**Sup. Figs. 27a-c** and **Sup. Note 9**), illustrating a challenge of utilizing naturally occurring orthologs due to requisite optimization and characterization to achieve efficient editing.
- Expanded discussions of related manuscripts: We have discussed previous manuscripts that have utilized machine learning for predicting Cas enzyme properties (**Sup. Note 12**). Notably, none of these studies generated predictions for non-canonical PAMs for their respective Cas enzymes, supporting the thesis of our work that our extensive training dataset comprised of enzymes with non-canonical PAM preferences is required to learn and predict PAM specificity.

Overall, we are grateful for the Reviewers' useful feedback that contributed to a greatly improved manuscript.

Referee #1

In this manuscript, Silverstein et al. described a framework of engineering PAM-altered SpCas9 proteins with help of machine learning (ML). By shortlisting functional candidates from a library of PI-domain-engineered SpCas9 variants and characterising their PAM preferences, the full mutational space was explored through ML, resulting in a model termed as PAMmla. While the ML method used here does not sound new, PAMmla identified various unique clusters of Cas variants with altered PAM selections that were more specific and efficient than the original training data. These novel variants were capable of editing various genomic loci with relatively lower off-target events while maintaining similar or higher activity level than reported PAM-relaxed Cas variants. By an in silico directed evolution (ISDE) approach, the full list of 64M SpCas9 variants included in PAMmla can be navigated with user-defined parameters with reduced computing power while still enabling selection of decently performed variants. By demonstrating the possibilities of predicting PAM-altered yet restricted Cas9 variants by PAMmla, the authors provided new insights in tailored Cas design for applications, in which PAM-relaxed variants were not usually favourable.

We thank Referee #1 for their positive assessment of our work and for appreciating the utility and novelty of PAMmla, and feel that our manuscript has been substantially strengthened by addressing their concerns.

Here are some comments and questions for the authors:

1. The authors suggested that they have "obtained 655 unique SpCas9 enzymes that survived the selections (Sup. Table 1)." (line 135). While there are 655 rows of entries labeled as "bacterial selections", only 634 of them were unique AA entries. For example, the enzyme "MQKSER" appeared 4 times, all labeled as "bacterial selections on NGCA PAM", but with slightly different values in the measurements. The authors should confirm the accuracy of the reported numbers and the table.

Response 1.1: We thank the Referee for noticing this. Several enzymes were recovered multiple times during different bacterial selections or included in separate HT-PAMDA experiments multiple times as controls (to assess inter-assay variability, which was minimal). Although we were careful to not include enzyme duplicates in the ML data set to avoid overlap between the training and test sets, we overlooked the duplicates while making **Fig. 1**. Our revised manuscript has now been updated to utilize only unique enzyme sequences; we report new numbers of unique enzymes in the text (634) and in **Figs. 1e** and **1f,** and have removed duplicate variants and re-analyzed **Sup. Fig. 5** (we note that the results for analyses were almost identical).

2. In the same table (Sup. Table 1), there are quite a lot of identical values in 14 decimal places (e.g. -8.46813474563119 appeared for 2752 times; -7.4827558925092 appeared for 3807 times etc.), are they some sort of defects or they are indeed the correct values to appear in those boxes? Overall, it is not very clear in how to interpret the table values.

Response 1.2: These numbers are the result of calculating $\log_{10}$ rate constants for PAMs where no cleavage activity was detected (fitting an exponential decay curve to an essentially flat line results in a very small rate constant approaching 0 – the specific numbers are likely an artifact of the limit of the curve fit algorithm used in the HT-PAMDA analysis pipeline). We determined that the detection limit of our HT-PAMDA assay is around 10^{-5} (-5 when looking at log10 values), thus rate constants lower than this are not meaningful and can be interpreted as having no detectable activity. For the purposes of our analysis and model training, all rate constants lower than 10^{-5} were set to 10^{-5} to avoid biasing the model with non-meaningful data as described in the methods section of our manuscript. Additional analysis of the detection limit of HT-PAMDA data is provided in **Sup. Note 5** and **Sup. Figs. 9h,l**.

3. While transformation efficiency and coverage of the 6AA_NNS library was not stated, "~24-48 colonies" (line 697) were selected by the authors from each selection for downstream characterisation using HT-PAMDA. It was however uncertain how representative 24-48 colonies can be in such context. This was essential to evaluate if the claims "an enzyme's most efficiently targeted PAM did not always correlate with the PAM on which that variant was selected (Fig. 1e, Sup. Figs. 4 and 5)." (line 203-204); "PAMmla resulting variants were more specific for the target PAM than those obtained from bacterial selections" (line 313-315) were resulted from sampling biases or the intrinsic issue of bacterial selection.

Response 1.3: We believe that an inability of bacterial selections to identify enzymes with high activity and high specificity and that are optimal on the PAM utilized in the selection is indeed an intrinsic 'bug' of the approach (and is also a bug of most evolution systems that do not implement a counter-selection). Most variants surviving the bacterial selection may not be specialized for the PAM of interest since the selection condition only requires activity on the PAM of interest to be above a certain threshold; the selected variants may therefore be more active on other PAMs that we were not selecting for (what we refer to as *relaxed* variants). We had previously speculated in our Discussion section and have modified this text:

"Prior to PAMmla, there were few examples of PAM-altered enzymes[1] relative to the many PAM-relaxed enzymes[2–5]. This discrepancy suggests that relaxing the PAM may be the simplest evolutionary trajectory to unlock new targeting capabilities for a Cas enzyme, perhaps because PAM-altered enzymes may require several specific simultaneous mutations that function epistatically to specify new PAMs. PAM-altered enzymes are therefore less likely to be discovered during

experimental engineering approaches (e.g. directed evolution) that do not incorporate a counter-selection step to preserve PAM selectivity."

Additionally, because the sample of bacterial selected enzymes that we characterized was considerably larger (24-48 enzymes per PAM) compared to the set of tested ML variants for each PAM (5-15 enzymes), we do not think that sampling bias affects our conclusions about PAMmla enzymes being more specific than those from bacterial selections. We most often recovered better enzymes from the PAMmla set despite being at a sampling disadvantage compared to the enzymes from bacterial selections. For clarity, we have added details regarding the SpCas9(6AA) library size and in our revised manuscript have included additional details about library construction and representation in the Methods section.

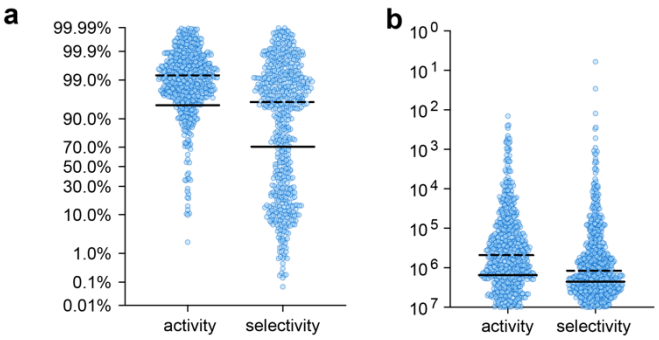
4. The authors mentioned that they compared different ML models, but the description of the ML methods used are very limited. Is this based on the published MLDE method, or any novel models were developed?

Response 1.4: We apologize for the lack of clarity and confusion, as we did not use the MLDE framework in our manuscript. Rather, we constructed our own neural network using TensorFlow. We have added additional details to the Methods section of our revised manuscript and a link to our GitHub page is now provided (https://github.com/RachelSilverstein/PAMmla).

5. Following from the authors' comments on the variants recovered from positive selection (line 311-313 and Fig 3f), the authors may want to indicate (in terms of activities ranking) where these positively selected variants are predicted to be among the full library (64 millions)? Are they at the top 10% range of the library? Should the authors expect to find more of the maximum efficiency enzymes predicted by PAMmla if they increase the stringency of the selection?

Response 1.5: We thank the Referee for this suggestion, which highlights the performance of PAMmla enzymes over those from the bacterial selections. We ranked the position of the enzymes derived from bacterial selections amongst all 64M SpCas9(6AA) PAMmla predictions, which revealed that the experimentally selected enzymes are on average within the top 10% of all PAMmla predictions in terms of activity (median within the top ~1%; **Sup. Fig. 13a**). Given our observations about PAM relaxation being the typical evolutionary trajectory during the selections, the PAM selectivity of enzymes from bacterial selections ranked on average in the top 70% of PAMmla predictions (**Sup. Fig. 13b**). We note that an enzyme in the top 10% is outperformed by 6.4M predicted enzymes. An inability to completely screen the large size of the 64M member library during the bacterial selections likely prevented us from recovering the most efficacious enzymes, highlighting the utility of identifying more optimal enzymes via PAMmla. Increasing the selection stringency may have recovered more active enzymes for some PAMs, but may have also resulted in higher activity but relaxed PAM variant enzymes (and thus would not have generated enzymes with improved selectivity).

Accordingly, we have added the following text to the manuscript: "*When ranked amongst PAMmla predictions of 64M enzymes, bacterial selection-derived enzymes ranked on average in the 95th percentile in terms of activity (outranked by 3.2M enzymes) and the 70th percentile in terms of selectivity (outranked by 18.9M enzymes) (**Sup. Figs. 13a,b**), highlighting the utility of PAMmla to predict superior enzymes versus those derived from experimental selections.*"



Supplementary Figure 13. Ranking of bacterial selection-derived enzymes amongst the SpCas9(6AA) library. *(**a**) Activity and selectivity rankings of SpCas9 PAM variant enzymes derived from bacterial selections ranked within the full set of 64 million PAMmla predicted enzymes. Rate constants (k) and selectivity (k/sum(all ks)) for the PAM on which each enzyme from bacterial selections was obtained were ranked amongst the distribution of activities/selectivities within the entire prediction set. Each datapoint represents the ranking of an enzyme derived from bacterial selections. The background distribution of all rate constants and selectivities for each of the 16 NGNN PAMs within the 6AA library was estimated by taking a random sample of 1 million enzymes from the library and generating PAMmla predictions. (**b**) Corresponding number of PAMmla predicted enzyme variants amongst the 64 million SpCas9(6AA) library that are predicted to exhibit improved activity or specificity compared to each enzyme in **panel a**. For **panels a** and **b**, solid lines represent means and dashed lines represent medians.*

6. Given the HT-PAMDA results of the selected variants from the bacteria screen, it was likely that variants could appear in multiple PAM groups and should naturally enriched more in PAMs that they edited more efficiently during bacteria selection. Labeling where these variants were selected according to the enrichment scores compared to the control plate would be a more sensible and fair way to depict an inferior position of bacteria screening approach.

Response 1.6: We were unable to generate enrichment scores for enzymes from bacterial selections, since the same enzyme was rarely recovered more than once from selections due to the high diversity of the library (enrichment of specific amino acids amongst the SpCas9(6AA) positions is plotted in **Sup. Fig. 2c**, rather than enrichment of entire enzyme sequences). We agree that the fairest comparison is with bacterial selected variants on the PAM for which they were selected, an analysis that is depicted in **Fig. 3f** (we note that the figure caption was incorrect from a previous iteration and

has now been updated) and also in **Sup. Fig. 5**. Our results reveal that few enzymes from the bacterial selection are most active or selective against the PAM that they were selected against, similar to as described in ***Response 1.5*** and in our revised manuscript in **Sup. Figs. 13a,b**.

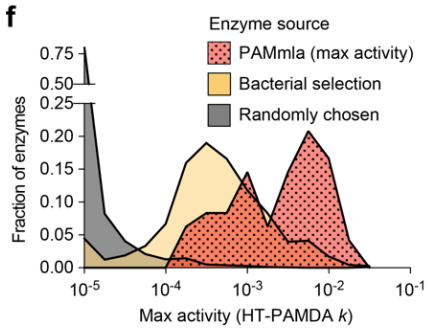
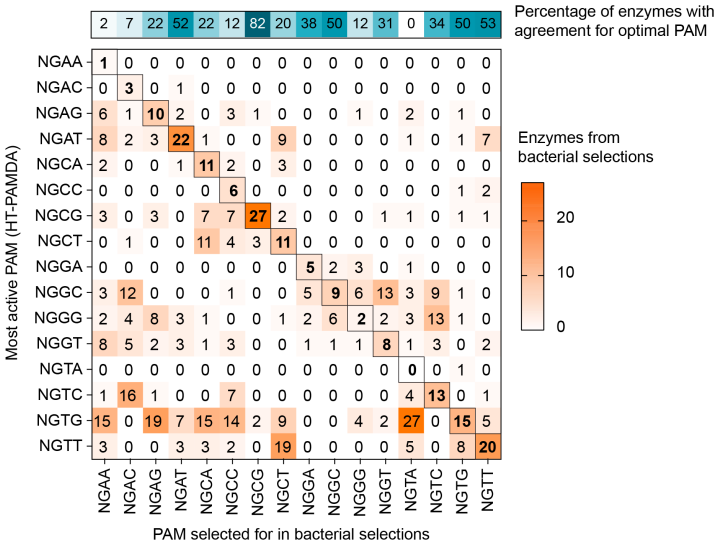


Figure 3. Characterization of the PAM requirements of PAMmla-predicted enzymes. *(**f**) Distribution of SpCas9 enzymes based on their experimental ks, with enzymes from 3 categories: randomly chosen from the SpCas9(6AA) library, derived from a bacterial selection, and predicted by PAMmla to maximize activity on one of the 16 NGNN PAMs. The plotted k is the rate constant on the PAM of interest (the PAM used in bacterial selection experiments, or the query PAM for PAMmla predictions). For PAMmla enzymes, the k plotted is from the PAM that we sought to maximize with PAMmla predictions.*



Supplementary Figure 5. Correlation between preferred PAM and sequence selected against. *Comparison of most preferred PAM for each enzyme (maximum k determined by HT-PAMDA) compared to the PAM on which each enzyme was selected for during the set of 16 bacterial-based selection experiments on NGNN PAMs. The outlined squares indicate the diagonal where there would be agreement between the most efficiently targeted PAM for the enzyme from HT-PAMDA data compared to the selection PAM.*

7. Only one of the spacer sequences used in HT-PAMDA was the same as the one used in bacteria selection. Would that set of data show more consistent result with the bacteria selection in terms of the most efficient PAMs? The authors should address the potential spacer discrepancy as well as the different environmental factors during bacteria screen and in vitro assay.

Response 1.7: When we mentioned discrepancies between HT-PAMDA profiles and an enzyme's preferred PAM resulting from bacterial selection experiments in our manuscript, we intended to refer to the intrinsic challenges of a bacterial selection rather than any deficiencies of HT-PAMDA to determine sufficiently accurate PAM profiles (for an extended discussion, please also see ***Response 1.3*** above). We have modified the text to try to make this point clearer.

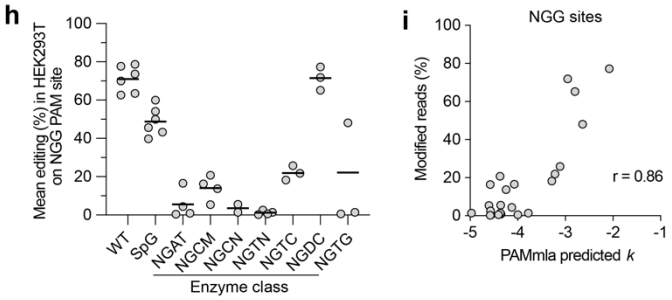
Data from both PAMDA spacers are in high agreement with one another (average Pearson r = 0.9; **Sup. Fig. 9h**), suggesting that the different spacers do not appear to be a major source of variation for HT-PAMDA results (similar to as originally investigated in our SpRY study; Walton et. al. *Science*, 2020). The PAM on which an enzyme was selected does typically exhibit some activity by HT-PAMDA (see **Fig. 3f**, above), supporting how enzymes can survive the bacterial selection with weak activity but that is sufficient to cleave the toxic plasmid encoding the modified PAM. For the small percentage of variants that survived a selection but were inactive on that PAM by HT-PAMDA, we believe these variants may be "cheaters" which survived the selection by some mechanism other than Cas9-based plasmid cleavage (possibly due to stochastic loss of the toxic plasmid or a mutation resulting in ccdB toxin resistance).

8. There are discrepancies on PAM specificity based on the bacterial selection versus the in vitro assay HT-PAMDA. It is not certain whether such discrepancies also exist when comparing HT-PAMDA and the assay results obtained from the experiments in mammalian cells. The authors should establish that correlation to confirm such consistency (similar to Fig. 3c, but with mammalian cell validation data). In mammalian cells, are the PAM-altered variants tested (for example those in Fig. 4b) not cutting the other PAMs (among the 16 NGNN PAMs) that they were predicted not to?

Response 1.8: Similar to as noted above in ***Response 1.3*** and ***Response 1.7***, discrepancies between HT-PAMDA profiles and an enzyme's preferred PAM are meant to reference the intrinsic challenges of a bacterial selection rather than any deficiencies of HT-PAMDA to determine sufficiently accurate PAM profiles. Our previous data supports that HT-PAMDA can

accurately characterize the PAM requirements of SpCas9 enzymes in human cells (see **Fig. S5f** from Walton et. al. *Science*, 2020). In our present study, we have not generated sufficient human cell data to confidently perform the requested comparison. We observed efficient human cell editing on the optimal PAMs for all ML-generated enzymes (as predicted by PAMmla and subsequently validated experimentally by HT-PAMDA) that we validated in HEK 293T cells (**Sup. Fig. 17a-g**). Unfortunately, it is not feasible to test all the enzymes in human cells at all of the 16 NGNN PAMs since such an experiment would involve thousands of arrayed transfections (and our validation experiment in human cells was already quite extensive with 40 enzymes and 48 genomic sites). However, to support this conclusion we did include a target site with NGG PAM for all enzymes to be used as a control (**Sup. Figs. 17a-g** and summarized in panels **h,i**). PAMmla enzymes predicted to be incapable of recognizing NGG PAMs led to only low-level editing on the control target site (specifically enzymes predicted to target NGAT, NGCM, NGCN, or NGTN PAMs). Other enzymes, such as those targeting NGDC (a degenerate PAM that includes NGG) were capable of editing at the control site with an NGG PAM. For NGTG enzymes, we observed only one enzyme MRRCQR could target NGG in human cells (**Sup. Fig. 17g**), but this was also apparent from our HT-PAMDA data (**Sup. Fig. 16**). Thus, our data supports that there is good agreement (r = 0.86) between whether an enzyme is predicted to target NGG and whether we observed editing on the NGG control site in human cells (**Sup. Fig. 17i**).

Future experiments to more thoroughly assess the performance of PAMmla-predicted enzymes in human cells will be informative, including to permit benchmarking of HT-PAMDA and/or PAMmla datasets. Such experiments will require more scalable assays, similar to those previously utilized to characterize SpCas9 PAM variants across thousands of target sites[6,7].



Supplementary Figure 17. Genome editing in human cells with PAMmla predicted enzymes. *(**h**) Summary of editing efficiencies for PAMmla predicted enzymes against the NGG PAM control site in HEK 293T cells tested in **panels a-g**. (**i**) Correlation between PAMmla predicted rate constant for the control NGG PAM tested for each enzyme in **panels a-g** versus the mean observed editing in HEK 293T cells at that PAM site. Enzymes predicted by PAMmla to have low activity on NGG PAMs showed low editing on NGG sites in human cells.*

9. Currently the off-targeting analysis of the MRRWMR enzyme was done in two separate sgRNAs instead of the one used in the RHO P23H locus, which it was designed to target. It would be helpful to also perform GUIDE-seq2 on this sgRNA to properly match the reason why this variant was selected in the first place.

Response 1.9: In our revised manuscript we have performed additional GUIDE-seq2 experiments to assess the specificity of MRRWMR, KRHWMR, SpG, and SpRY when paired with the *RHO* P23H gRNA (new results now presented in **Fig. 6g** and **Sup. Figs. 28d-f**). Briefly, as we anticipated with PAMmla enzymes predicted to be PAM-selective, we observed much higher specificity with MRRWMR and KRHWMR compared to SpG and SpRY via a higher fraction of GUIDE-seq2 reads attributable to the on-target site (**Fig. 6g**). Furthermore, with the PAMmla enzymes we observed substantially fewer off-target sites detected by GUIDE-seq2 (**Sup. Figs. 28e-f**), despite observing similar levels of on-target dsODN incorporation in the experiments that suggests comparable overall efficiency (**Sup. Fig. 28d**).

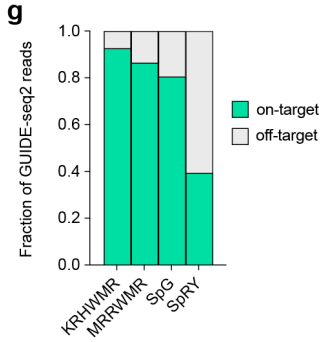


Figure 6g. *Fraction of GUIDE-seq2 reads attributed to on- and off-target sites for PAMmla predicted enzymes, SpG, and SpRY when paired with the RHO P23H sgRNA in homozygous P23H HEK 293T cells (see also **Sup. Figs. 28d,e**).*

10. In Figure 6g, while “reads containing indels that span the P23H mutation, edited counts were distributed using the ratio of WT to mutant as observed for the identifiable edited reads”, the authors should also indicate how much of the edits contained such indels for better evaluation of the data. The assumption to distribute them in ratio of WT to mutant may not necessarily correct but the effect would be negligible if such event was rare to start with.

Response 1.10: In our revised manuscript we have now included this data in **Sup. Figs. 26d** and **29a** (for experiments in human cells or mice, respectively). For experiments in HEK 293T cells, the percentages of unidentifiable reads were roughly proportional to total editing (as expected since unidentifiable reads arise from large mutations spanning the PAM). For *in vivo* experiments, the unidentifiable reads were much less common potentially due to a lower propensity for larger deletions.

11. It would be great if the reason why the KRHWMR enzyme can specifically edit NGTG but not NGGG as demonstrated in Figure 6g can be speculated through structure modeling. This could further strengthen the idea of allele-specific editor design using the PAMmla and provide more insight on PAM-altered Cas design.

Response 1.11: We thank the Referee for this comment and have added many new analyses related to structural predictions of PAMmla enzymes to understand their modified PAM interactions. Our revised manuscript includes SHAP analyses, structural modeling, and a substantial discussion of potential functional roles of amino acid substitutions in PAMmla enzymes (see page 13 in the manuscript and new **Sup. Note 6**, **Sup. Figs. 14** and **15a-l**).

For MRRWMR and KRHWMR and their specific preferences for NGTG over NGGG, we performed SHAP analysis and structural modeling, now described on page 20 of our manuscript with the following text:

"*To examine potential structural determinants of the PAM-selective preferences of MRRWMR and KRHWMR, we modeled their mutations and investigated their contributions to PAMmla predicted rates on NGTG and NGGG PAMs using SHAP (**Figs. 6h-k**, **Sup. Figs. 30a-e** and **Sup. Note 10**). For both enzymes, PAM selectivity for a 3rd PAM position T over G largely resulted from the E1219W and R1335M side chains forming a hydrophobic pocket, enabling discrimination via favorable hydrophobic interactions with the T3 methyl group of an NGTG PAM and unfavorable interaction with the polar G3 side chain of NGGG (**Figs. 6h,i**). SHAP analysis supports that these mutations positively impact recognition of NGTG and negatively impact activity on NGGG (**Figs. 6j,k** and **Sup. Figs. 30b-e**). Additional common mutations between the two PAMmla enzymes S1136R and T1337R contribute to PAM selectivity and/or potentiation of activity (**Sup. Figs 30a-e** and **Sup. Note 10**).*"

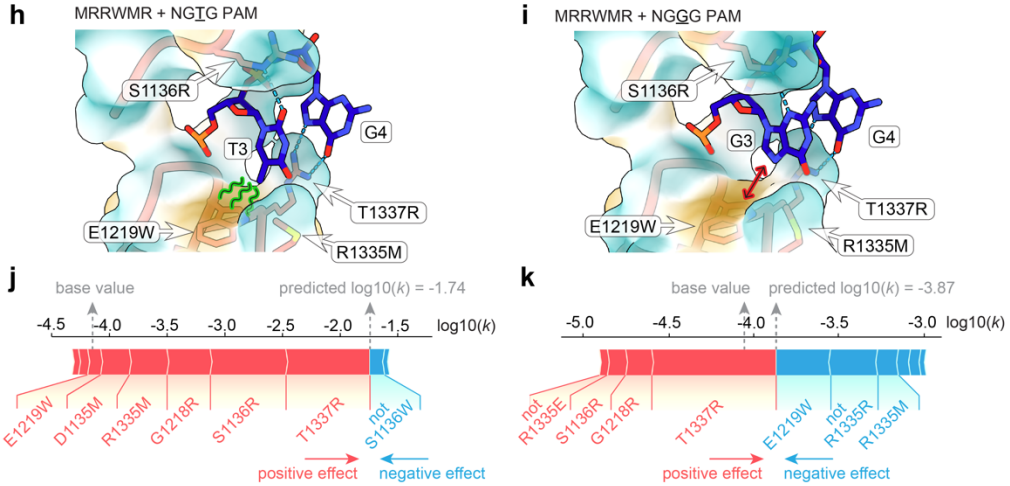
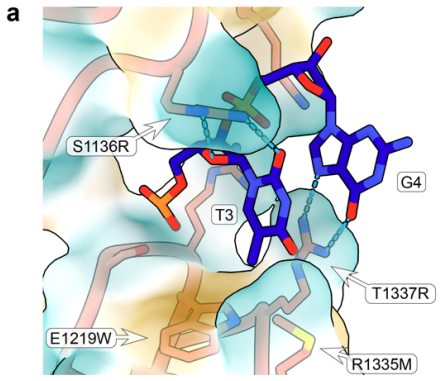


Figure 6. In silico directed evolution of an allele-specific editor for the RHO P23H allele. *(**h, i**) MRRWMR modelled on the structure of VRER (PDB: 5FW3)[8] interacting with NGTG or NGGG PAMs (**panels h** and **i**, respectively). Protein surface is colored by lipophilicity potential. Hydrogen bonds are represented by dashed lines and Van der Waals interactions are represented by green squiggles. (**j, k**) Force plots depicting SHapely Additive exPlanations[9] (SHAP) values for MRRWMR activity on NGTG or NGGG PAMs (**panels j** and **k**, respectively).*

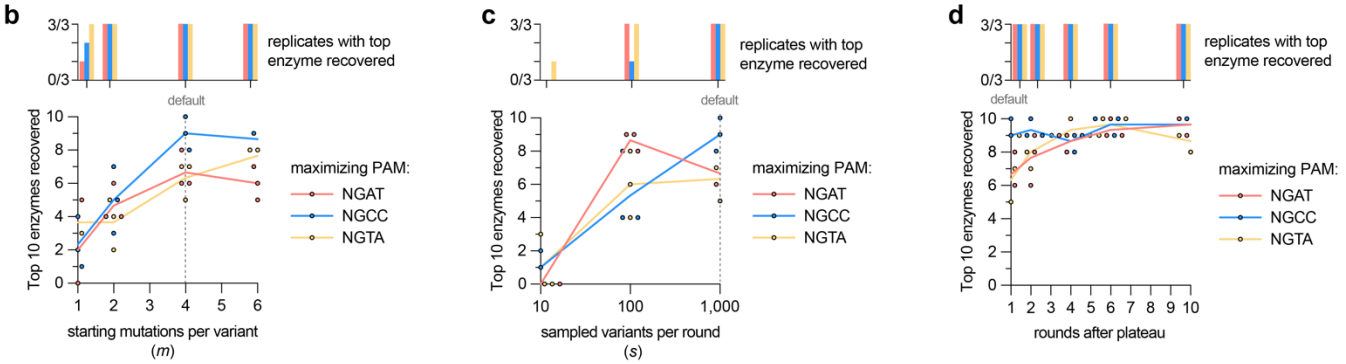


Supplementary Figure 30. Analysis of factors contributing to MRRWMR and KRHWMR PAM preferences. *(**a**) Structural prediction of an alternative conformation of the S1136R mutation leading to additional hydrogen bonding with T at position 3 of the PAM.*

12. The authors suggested the use of ISDE over direct sorting approach when handling the 64M prediction list due to its lower computing demand yet still being able to select desired variants. The authors should benchmark the two methods in these two aspects to demonstrate that ISDE was indeed faster to work with and capable of selecting the variants comparable to the exhaustive sorting approach. This should further demonstrate that ISDE was the more favourable option.

Response 1.12: To address the Referee's comment, but also balanced by comments from other Referees to minimize the ISDE aspect of our work, we have added additional text to **Sup. Note 8** to better describe the use and performance of PAMmla-ISDE. The relevant paragraph from **Sup. Note 8** and panels from **Sup. Figs. 23b-d** are included below:

*"Although ISDE permits rapid and interactive enzyme prioritization, it does not ensure convergence on the most optimal predicted enzyme and carries the risk of landing in local maximums of the fitness landscape (similar to traditional experimental directed evolution). This risk can be mitigated by repeating the ISDE process from different starting sequences and varying ISDE parameters. To explore the effectiveness of ISDE at reaching the global fitness maximum, we compared enzymes obtained by exhaustive sorting to those obtained through ISDE for activity maximization on three PAMs (**Sup. Figs. 23b-d**). When using ISDE with default parameters (available at https://pammla.streamlit.app/Evolve_Variants), we recovered the true maximum enzyme for all 3 replicates for each of the 3 PAMs tested. Of the top 10 true maximum enzymes, we recovered between 5-10 depending on the PAM and replicate. We tested the effect of varying ISDE parameters and found that increasing the number of allowed mutations per enzyme (**Sup. Fig. 23b**) and sampling a greater number of mutants per round increased the probability of recovering top variants (**Sup. Fig. 23c**). Inclusion of additional rounds of evolution following fitness plateau increased the proportion of the top 10 variants recovered (**Sup. Fig. 23d**). In general, although a single ISDE run with default parameters should be sufficient to obtain the true maximum enzyme, additional rounds of evolution and mutations per variant can be utilized to increase the proportion of top enzymes recovered. Performing 3+ repetitions of ISDE and compiling results may be desirable to predict a larger number of maximally optimal enzymes. While it is currently possible to generate predictions for all 64 million possible enzymes from the PAMmla model, if the model is expanded to include additional amino acid positions, the possible sequence space becomes too large to generate all possible predictions. Therefore, methods such as ISDE will become necessary, where computation time does not depend on the size of the sequence space being explored."*



***Supplementary Figure 23. Design and validation of in silico directed evolution**. (**b-d**) Effect of ISDE parameter values on the identification of optimized PAMmla predicted enzymes, including varying the number of starting mutations per round (m) (**panel b**), random variants generated per round (**panel c**) and number of additional evolution rounds performed once a plateau is reached before decreasing m (**panel d**). Proof-of-concept PAMmla-ISDE runs were performed to identify enzymes with maximal activity against NGAT, NGCC, or NGTA PAMs. Aside from the parameter being tested, ISDE was run with default parameters of 1,000 random starting sequences, m = 4 starting mutations per enzyme, s = 1,000 sampled enzymes per round, n = 10 top variants to keep per round, and p = 1 additional round of evolution after a plateau is reached. The number of true top 10 predicted enzymes, determined by exhaustive sorting of PAMmla predictions, recovered by ISDE are shown. Top bar graphs represent the number of replicates in which the most optimal enzyme was recovered.*

13. Figure 6 c-f indicated that the ISDE regime would identify enzymes with quite different maximum activities in each trajectory. In general, how many ISDE runs the authors would recommend to ensure that one may find the optimal variant for experimental application? Will there be different rounds of ISDE iteration based on the position of the SNP on the PAM?

Response 1.13: We have added additional recommendations for use of ISDE in **Sup. Note 8** (and please see ***Response 1.12*** above). Briefly, we recommend that: "*In general, although a single ISDE run with default parameters should be sufficient to obtain the true maximum enzyme, additional rounds of evolution and mutations per variant can be utilized to increase the proportion of top enzymes recovered. Performing 3+ repetitions of ISDE and compiling results may be desirable to predict a larger number of maximally optimal enzymes*".

14. Supplementary Note 5 claims that active variants typically have PAM activity correlations of 0.7 to 0.9. The authors should elaborate on that and plot such correlations among PAM sequences. The authors should also elaborate about the definition of "inactive enzymes" here, i.e. how many inactive PAM sequences (cut off k > 10^-4?) an inactive enzyme has to have in order to achieve such correlation of 0.1?

Response 1.14: Plots for correlation between HT-PAMDA replicates and PAMmla predictions for inactive enzymes are provided in **Sup. Figs. 9h,i**; we apologize and have corrected this error, as the figure references in **Sup. Note 5** in our original submission incorrectly referenced **Sup. Figs. 7g,h**).

Minor questions:

15. Fig 3c and S9b, h, and g: Are the predicted and empirical k values derived from 1 specific PAM sequence or averaged across many PAM sequences?

Response 1.15: Each predicted and empirical datapoint is a single rate constant on one PAM. Each enzyme will have 64 data points on the graph (one for each NNNN PAM); this point has now been clarified in figure legends.

16. Figure S9d and f: Can the authors explain on how the breakdown values were calculated?

Response 1.16: Each enzyme is assigned a most preferred PAM based on empirical HT-PAMDA rate constant. An enzyme's nucleotide preference at a certain PAM position is defined as the nucleotide at that position in the enzyme's most preferred PAM. We have now added this explanation to the figure legend.

17. The detailed GUIDE-seq2 data were currently not available for most of the GUIDE-seq2 experiments. Please provide them either as supplementary tables or visualised figures as shown in Supp. Figure 19b.

Response 1.17: We have now provided all GUIDE-seq2 data outputs as **Sup. Table 6** and have included some GUIDE-seq2 data visualizations in **Sup. Figs. 22b** and **28e**; other visualization outputs were too large to include as **Sup. Figs**.

18. Supplementary Note 5: Where are Sup. Fig. 7g and 7h?

Response 1.18: Thanks! This was a typo that has now been corrected to refer to **Sup. Figs. 9h,i**.

19. Single guide RNA (sgRNA) might be a more proper term than just guide RNA (gRNA) as the latter was used to describe the crRNA + tracrRNA complex, which was not linked by a linker as the modern sgRNA that was used in this study.

Response 1.19: We have changed gRNA to sgRNA to accommodate the Referee's request.

Referee #1 (Remarks on code availability):

There is no supporting github codes provided in the article. The ML seems to based on the published MLDE framework (Wittmann et al., Cell Syst., 2021) (i.e., it is not very clear as only very limited description on the ML method used is provided). If so, such MLDE framework for CRISPR engineering was previously demonstrated (Thean et al., Nature Comm, 2022), which is also cited by the authors.

Response 1.20: The work in our manuscript does not utilize the MLDE framework (see ***Response 1.4*** above). We apologize for lack of details in our original submission and have added additional descriptions to the Methods section of our revised manuscript and a link to our GitHub page (https://github.com/RachelSilverstein/PAMmla).

Referee #2:

SpCas9 is the most popular gene editing tool and numerous approaches have been utilized to engineer this protein for optimal target selectivity and specificity. One key factor is its requirement for the specific detection of a PAM sequence which limits the range of available target sites. Previous research focused on the engineering of the PAM interacting (PI) domain, resulting in commonly used Cas9 variants with relaxed PAM recognition or altered PAM requirements. In the present manuscript, Silverstein et al. built on this concept, assayed a saturation mutagenesis library and used this data to train a neural network to identify novel Cas9 variants with unique PAM preferences. Successful examples of this approach are described as selected Cas9 variants were identified to exhibit novel PAM preferences.

In general, this is a highly relevant topic with a lot of competing research activities. Machine learning was for example used to optimize sgRNA selection (10.1016/j.celrep.2024.113765), but also applied for the detection of novel Cas9 PAM-interacting domains (doi.org/10.1371/journal.pcbi.1011621). Alternative approaches to acquire a larger selection of specific PAM requirements include the screening of natural Cas9 variants from other host bacteria or the generation of Cas9 variants without PAM requirements. The authors provide convincing arguments against the latter approach.

We are grateful to the Referee for their positive remarks and thoughtful feedback on our manuscript. Our revised paper reflects their recommended changes, for which we believe our resubmission is now considerably improved.

The focus of this manuscript are six amino acids in the PI-domain of SpCas9 that were subjected to saturation mutagenesis. These residues were specifically selected as their alterations was previously shown to alter PAM specificity. Here, the authors characterized the PAM requirements of SpCas9 enzymes with variant PI domains using HT-PAMDA. The investigated PAM is (N)GNN and I initially wondered why the G at the second position was fixed for the training set as (N)NN would also result in 16 screenable experiments. This is detailed in supplementary note 3, but I would find it important to indicate in the main text that the second position could not be changed due to its interactions with a network of amino acids coordinated by R1333. In theory, being able to change the second PAM position would be more desirable than changing only the third and a "new" fourth position. The rationale for this PAM library should also be shortly discussed outside of the supplementary note.

The bacterial-based selection approach provided 655 Cas9 variants with unique PI domains and revealed that most enzymes utilized a canonical NGG PAM or a relaxed PAM recognition, indicating that the PI domain co-evolved with PAM elements using regions outside of the selected six amino acids. I agree with authors' discussion point that it would be necessary to expand the six amino acid library in the future.

Response 2.1: We appreciate the comment and agree that varying the 2nd position of the PAM could also be useful. Our initial experiments revealed that modification of R1333 (which contacts the 2nd guanine of the PAM) typically must be accompanied by other compensatory mutations in poorly understood neighboring amino acids. We agree with the Referee that such experiments would be quite interesting to pursue in future work. Accordingly, we have added the following explanation to the main text (in addition to the expanded discussion in **Sup. Note 3**):

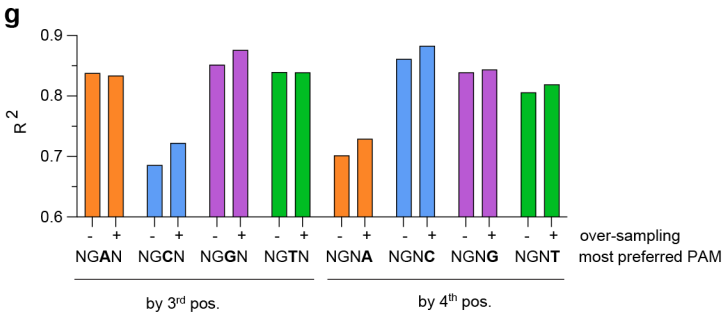
*"We selected residues D1135, S1136, G1218, E1219, R1335, and T1337 for saturation mutagenesis (leading to an SpCas9(6AA) library of plasmids). These six amino acids modulate specificity for the 3rd and 4th PAM bases in the context of the engineered enzymes SpCas9-VRER, SpCas9-VRQR and SpG[1,2,10] (**Sup. Note 3**). Expansion of our library to also mutate R1333, the residue which contacts the second guanine of the SpCas9 PAM[11], largely abrogated enzyme activity (**Sup. Note 3**). We therefore focused our efforts on altering the 3rd and 4th positions of the PAM."*

As 18% of 135 randomly mutated versions (without selection) provided (non-canonical) PAM preferences, the authors then selected a machine learning approach termed PAMmla to predict quantitative rate constants (ks) for Cas9 variations for each of the 64 (N9)GNN PAM versions. I am not an expert on machine learning and will refrain from reviewing this portion of the manuscript in detail. However, I would be curious how one measures that the training input was of sufficient quality. Most enzyme variants (82%, according to the 135 randomized versions) should result in versions that are not capable of editing any of the investigated PAM sequences. Should an ideal training set then not include 82% of inactive variants to better evaluate amino acid combinations that result in non-functional enzymes? What happened if this correction approach was not applied: "To account for the imbalanced nature of our training data towards NGGN and NGTN enzymes and those preferring a G at the 4th 237 PAM position (Fig. 1f), we assigned a label to each training example based on its most active PAM and randomly over-sampled to balance across PAM classes"?

Response 2.2: Our results suggest that the quantity, rather than proportion, of active and inactive enzymes used in our training set was of sufficient quality for the model to learn features of what constitutes an inactive enzyme (based on our evaluation on the test set; see **Figs. 2d,e**). We agree with the Referee that the inclusion of additional data for inactive enzymes would have been a logical and important next step had our model not performed as well on this training set composition. We hypothesize that the model can perform well with relatively few training examples for inactive enzymes because the most informative enzymes are those with altered PAM profiles (harboring meaningful data to train on).

Additionally, we were most interested in learning what predicts active rather than inactive enzymes, hence the skew in our training data; however, the performance of PAMmla suggests that our model can accurately predict both classes of enzymes.

Regarding our random over-sampling approach, we have included a new analysis to investigate the importance of over-sampling (**Sup Fig. 9g** in our revised manuscript). Although random over-sampling does not dramatically alter overall performance of the model, it slightly improved performance on PAM classes that were underrepresented in the training set.



Supplementary Figure 9. Machine learning models to predict PAM profile from amino acid sequence. *(**g**) Effect of random over-sampling by most active PAM. The PAMmla model was trained with and without randomly over-sampling the training set to balance the number of enzyme variants with different PAM preferences. R^2 values for the two models were compared on subsets of variants within the test set with different preferences at the 3rd and 4th positions of the PAM. Over-sampling improved performance particularly for under-represented PAM classes.*

The authors determined PAM profiles for 253 novel Cas9 versions and verified editing efficiencies for PAM elements that have not been described before. In general, the performance of PAMmla is impressive and predicted PAM requirement could be confirmed. Interestingly, the generated enzymes usually showed specific PAM requirements and not only relaxed PAM recognition. I am wondering if structure modelling is able to explain this feature. Here, it would be desirable to analyse the modelled structure of PI domain variants (...for each of the HT-PAMDA profile clusters) and dock DNA targets with specific PAM versions. This approach should help to validate the authors' results and also stimulate further structure-based engineering of amino acids near the six mutated residues.

Response 2.3: We thank the Referee for this comment and agree that given the large catalog of SpCas9 variant enzymes that we have characterized (1,000+ enzymes either through bacterial selections or PAMmla predictions), it is interesting to speculate about potential mechanisms for PAM interaction. Accordingly, we have added many new analyses related to structural predictions of PAMmla enzymes to understand their modified PAM interactions. Our revised manuscript includes SHAP analyses, structural modeling, and a substantial discussion of potential functional roles of amino acid substitutions in PAMmla enzymes (see page 13 in the manuscript and new **Sup. Note 6**, **Sup. Figs. 14** and **15a-l**). Furthermore, we performed similar analyses for specific enzymes described in our *RHO* P23 experiments (MRRWMR and KRHWMR) and their specific preferences for NGTG over NGGG. For these enzymes we performed SHAP analyses and structural modeling that are now described on page 20 of our manuscript and in **Figs. 6h-k**, **Sup. Figs. 30a-e,** and **Sup. Note 10**. Although it would have been interesting to perform these same analyses for all potential new clusters of PAM variants, we found ourselves struggling to decide which to pursue (given the many ways that enzymes can be prioritized based on activity or specificity). In future studies we agree that it would be interesting to more thoroughly elucidate the functional roles of amino acid substitutions for more diverse PAM classes of PAMmla enzymes.

The manuscript is written in clear and coherent fashion and I only disliked that crucial information is "hidden" in eight supplementary notes.

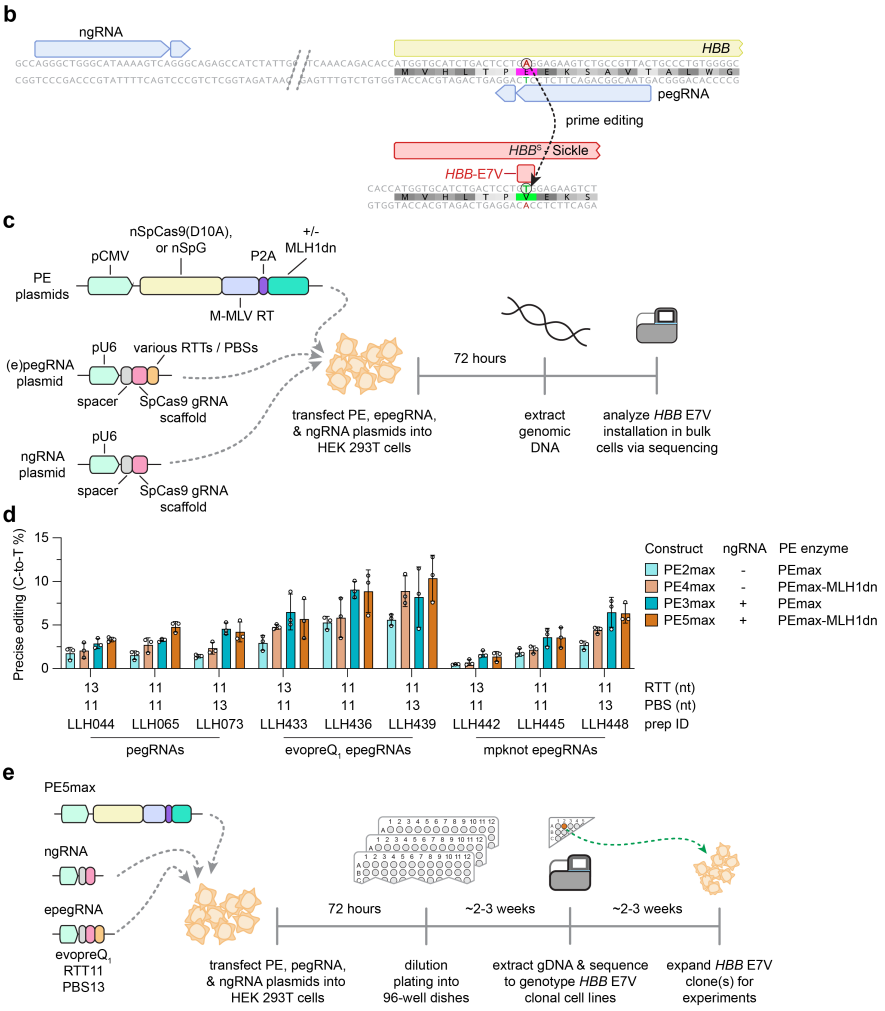
Response 2.4: We are gratified to hear that the Referee enjoyed reading our manuscript, and agree that the word limit constraints unfortunately prohibit inclusion of many interesting details in the main text of our manuscript (and are already 2x the recommended work limit in the main text!). We hope that the text in the Supplementary Notes is accessible to the Referee and all readers.

A reference to this paper (Malbranke, C et al. (2023) PLoS Comput Biol 19(11): e1011621) would be appropriate to indicate an alternative ML approach to generate Cas9 variants with modified PAM specificity.

Response 2.5: Thanks for this comment! We have added this citation (and a citation for another recent preprint) to our revised manuscript and have briefly mentioned how these studies relate to our current work: "*previous ML-based approaches have sought to predict sequence-diversified but functional Cas9 enzymes[12,13] or to modify alternate Cas properties like the on-target activity of SaCas9-KKH[14,15] (**Sup. Note 12**). While these studies support the use of ML for Cas enzyme engineering, they also demonstrate that natural sequence data is likely insufficient to train models to predict SpCas9-based enzymes with altered PAM requirements. The use of experimental data to train PAMmla overcomes this limitation by producing hundreds of non-natural SpCas9 proteins with properties not yet identified in nature.*"

Few references need to be updated to be able to follow the methods section and the data availability section will need to be adjusted when primary data is available:
- Line 821: As described elsewhere (Hille LT, Christie KA, Walton RT, et al., submitted)

Response 2.6: We have removed this reference to our other manuscript and have instead included the relevant primary data in our revised manuscript (see descriptions of generating the *HBB* E7V cell line on page 16 of the text, page 31 in the Methods section, and an updated **Sup. Fig. 20** with new figure panels **20b-e**, shown below).



Supplementary Figure 20. Creation of a sickle cell HBB E7V cell line and base editing with ML-derived PAM variants (***b***) *Schematic of a region of the human HBB gene surrounding amino acid E7. To install the HBB E7V mutation, prime editing guide RNA (pegRNA) target sites and nicking sgRNA target sites (ngRNAs) were designed for compatibility with NGG PAMs. (**c**) Workflow for assessing the editing efficiencies of different prime editor (PE) constructs, pegRNAs, and ngRNAs to generate the HBB E7V mutation in HEK 293T cells in bulk transfection experiments. (**d**) Precise prime editing efficiencies using to generate the HBB E7V mutation in HEK 293T cells. Editing efficiencies were assessed by targeted amplicon sequencing and modified reads were analyzed using CRISPResso2; mean, standard deviation, and individual datapoints shown for n = 3 independent biological replicates. (**e**) Workflow for generating a clonal HEK 293T cell line harboring the HBB E7V mutation using PE4max, the evopreQ$_1$ RTT11-PBS13 epegRNA, and ngRNA.*

- line 867: GUIDE-seq2 reactions were performed essentially as described (Lazzarotto & Li et al, in preparation) with minor modifications)

Response 2.7: The manuscript describing GUIDE-seq2 is currently submitted to another journal by Dr. Shengdar Tsai's laboratory. We will discuss with the editorial team what the best approach is for citing this work, and will confer with Dr. Tsai about his plans to preprint their manuscript (once he returns from paternity leave!). In the meantime, here is a link to a recording where Dr. Tsai briefly describes GUIDE-seq2 in a recorded talk for the FDA (from December 2022):
https://www.fda.gov/vaccines-blood-biologics/workshops-meetings-conferences-biologics/assessing-genetic-heterogeneity-context-genome-editing-targets-gene-therapy-products-12162022
Please click on "Genome Editing Workshop, Recording 1" and scroll to the 2:36:30 timepoint.

- line 889: a summary of GUIDE-seq2 data will be made available in Sup. Table 6

Response 2.8: We thank the Referee for noticing this omission and have included this Supplementary Table (**Sup. Table 6**) along with our revised manuscript.

- several instances: (to be deposited on Github).

Response 2.9: We have provided a link to our repository on Github (https://github.com/RachelSilverstein/PAMmla) and these details have been updated in our revised manuscript.

Referee #3:

In this work the authors gather a unique dataset on the activity of Cas9 variants on different PAM sequences. A mutant library with randomized amino acids at 6 important positions was generated and first screened using a bacterial selection system on each of the 16 possible NGNN PAMs. 655 Cas9 variants were then cloned and their PAM specificity profiled using the HT-PAMDA method. Interesting variants were selected which provided insights into PAM binding specificity determinants, but the process was not satisfactory in terms of isolating efficient and specific variants against non-canonical PAMs.

The authors then generated additional data on 135 randomly chosen variants and used the total dataset to train models able predict PAM specificity profiles based on the 6 residues modified in the library. Their PAMmla model achieves successful predictions of PAM profiles and was then used in the design of variants able to target specific loci efficiently. 253 active variants were generated and their activity measured. The authors further successfully test PAMmla predicted nucleases and base editors in human cells, showing activity higher than the relaxed SpG-Cas9 protein as well as more specificity. Finally the authors generate variants with the goal to discriminate between two closely related PAM sites. Here too the model is successful at identifying a variant specific for the desired PAM.

Altogether the authors generated a very impressive amount of high-quality data which provides important novel insights into the determinants of Cas9 PAM specificity. They convincingly show that this novel dataset can be used to train models that are useful in the design of novel Cas9 variants. While I applaud the authors for the quality of the experimental work, which will be an important ressource for the community, I have a few important criticisms.

Response: We thank the Referee for their positive feedback, helpful suggestions, and thoughts on the applicability of PAMmla. Our manuscript has been substantially improved owing to their feedback.

The authors demonstrate improved activity versus the PAM-relaxed enzyme SpG, but not against natural Cas9 variants. Is there still a margin of improvement here in term of activity and specificity? How does the engineering of SpCas9 variants compare to the use of natural Cas9 variants with different PAM specificities or the construction of chimeric Cas9 variants that swap the PI domain entirely with that of natural variants?

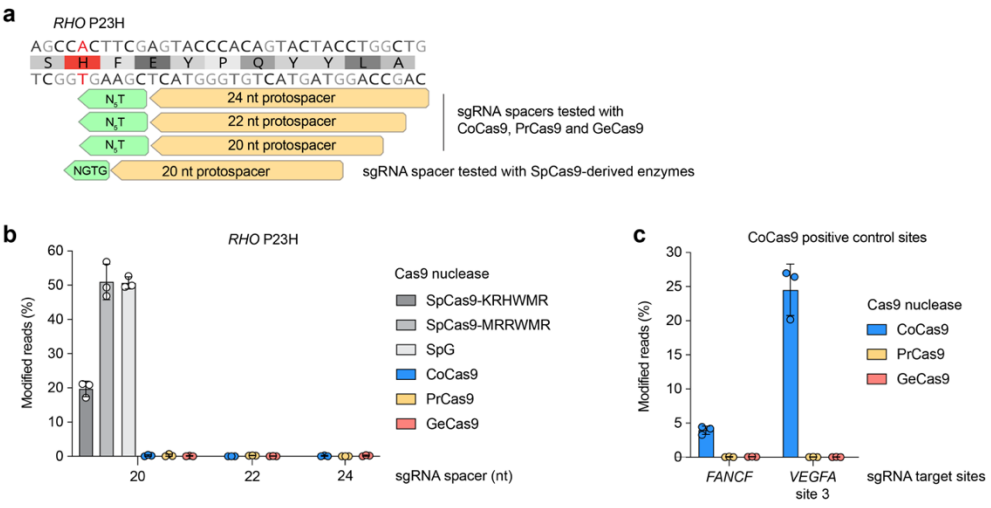
Response 3.1: Natural Cas9 variants have a range of reported PAM specificities. However, to the best of our knowledge, no other Cas9 orthologs (or Cas12a, etc.,) have been reported that reproducibly have comparable on-target editing to SpCas9. This was one of the main motivations for our study – to leverage key properties of SpCas9's high activity and high specificity, while overcoming the targeting range limitations of NGG by reprogramming a single well-studied protein scaffold to recognize new non-NGG PAMs. When utilizing alternate Cas orthologs, there is substantial requisite characterization and optimization of the enzyme and sgRNA; these new systems may have different properties in terms of genome wide specificity, cleavage profile, sgRNA spacer length, and optimal guide scaffold configurations, etc. Thus, although Cas orthologs and chimeric enzymes certainly have notable use cases, leveraging the broader diversity of these tools imposes many challenges for use. Simply put, they would not be as 'plug-and-play' into established SpCas9-based workflows in most laboratories. Conversely, SpCas9-derived PAMmla enzymes are simple to implement and utilize, given that there are only minor differences to the protein and most other characteristics for use remain constant.

To address the Referee's comment, we were intrigued to compare PAMmla-predicted enzymes against recently reported Cas9 orthologs. In the revised version of our manuscript, we now include new datasets to directly ask this question by testing naturally occurring Cas9 orthologs that have reported to be compatible with PAMs for *RHO* P23H targeting (including PrCas9[16], CoCas9[17], and GeCas9[17]). Unfortunately, we were unable to detect editing against in a homozygous P23H cell line with any of these 3 Cas9 orthologs, despite observing some editing with CoCas9 at two previously reported positive control sites (see new data in **Sup. Figs. 27a-c**). We describe this result in the text: *"Use of other recently described Cas9 ortholog nucleases PrCas9[16], CoCas9[17], and GeCas9[17] harboring PAM preferences that should target the P23H site failed to elicit detectable editing at the RHO locus (**Sup. Figs. 27a-c** and **Sup. Note 9**)."*

Please also see **Sup. Note 9** for an extended discussion of these results. Our comparison illustrates a challenge of utilizing Cas orthologs, especially when they have not been extensively characterized in previous studies. For instance, the enzymes that we tested in our revision may have uncharacterized target site preferences that differ from SpCas9, limiting which target sites they are compatible with (i.e. because of spacer lengths, spacer preferences, nuanced PAM preferences, etc.). Both GeCas9 and PrCas9 may also not be compatible with targeting the human genome, as previous studies have only demonstrated *in vitro* or plasmid targeting in human cells. As has been shown previously for other Cas9 orthologs, challenges for GeCas9 and PrCas9 might include weaker DNA binding[18], a reduced ability to target chromatinized genomic DNA[19], or a requirement for higher magnesium concentrations than what is available in human cells compared to bacteria[20].

Chimeric Cas9 enzymes retain the PAM preference from the PI domain of the original enzyme, but often with weaker activities. While the use of chimeric Cas9 enzymes has shown promise, this approach would not be as modular as PAMmla

enzymes (wherein only a few amino acid changes are necessary) and would be limited to the naturally occurring PAM preferences of efficacious orthologs whose PI domains are compatible with being grafted on to the remainder of SpCas9.



Supplementary Figure 27. Assessment of Cas9 orthologs for allele-specific targeting of RHO P23H*. (**a**) Schematic of RHO P23H target sites. (**b**) Nuclease-mediated genome editing at the RHO P23H locus in homozygous P23H HEK 293T cells, utilizing sgRNA and enzyme expression plasmids for PAMmla-predicted SpCas9-KRHWMR and -MRRWMR, SpG, CoCas9[17] (N_4GT PAM), GeCas9[17] (N_4GT PAM), and PrCas9[16] (N_5T PAM). sgRNAs encoding three different spacer lengths were tested for the non-SpCas9 enzymes. (**c**) Editing at previously described positive control genomic sites for CoCas9[17] (for PrCas9 and GeCas9, the same spacer sequences were used but combined with the ortholog-specific scaffold sequences[17,16]. Editing efficiencies in **panels b,c** were assessed by targeted amplicon sequencing and modified reads were analyzed using CRISPResso2; mean, standard deviation, and individual datapoints shown for n = 3 technical replicates.*

Importantly, I find the modeling aspects of the manuscript lacking. This is concerning given that the angle taken by the authors in their writing is putting the machine learning part forefront. The modeling methods are not described in sufficient details and could be performed to a higher standard. How were the model hyperparameters chosen? Did the authors conduct an hyperparameter search? The network used is already fairly small, but could an even smaller model work as well, in particular considering that the size of the training set is fairly small? On that note, what was the total size of the dataset used? Is it 655 + 135 ? It should be much easier to find this information in the manuscript.

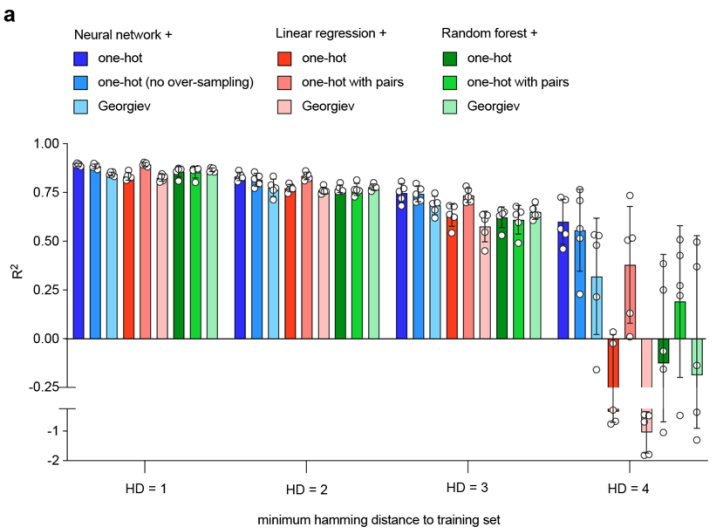
Response 3.2: We apologize for the incomplete details and have added additional text to the Methods section of our revised manuscript (including explicitly stating that the model was developed using HT-PAMDA data from 634 unique enzymes from the bacterial selection, and randomly chosen enzymes from the SpCas9(6AA) library). While testing various models, we found that the exact model size and hyperparameters chosen did not greatly impact the results. Rather, we believe that a main point of novelty of this paper is the data on which the model was trained, which allowed us to generate highly accurate predictions and many novel enzymes which will enable new capabilities using a simple modeling approach. We have edited the text of the manuscript to try to more accurately convey how certain models were selected and to better represent the main points of novelty of our work. Whether smaller models would be similarly functional given our training set size could be interesting to investigate in future studies.

The authors perform some comparison between their neural network model, a linear model and a random forest model, however it seems like no effort was made to encode interaction features in the linear model. Encoding all pairwise interactions and fitting a linear model with regularization is likely to provide similar performance to the neural network model while being a lot easier to interpret. Such feature encoding could also benefit decision tree based models, for which methods such as SHAP (SHapley Additive exPlanations) can provide interpretable insights.
Could the authors also perform some analysis of the features captured by the model to predict active Cas9 enzymes?

Response 3.3: Thank you for these helpful comments! We have performed additional analyses related to these comments and feel the results have strengthened our manuscript.

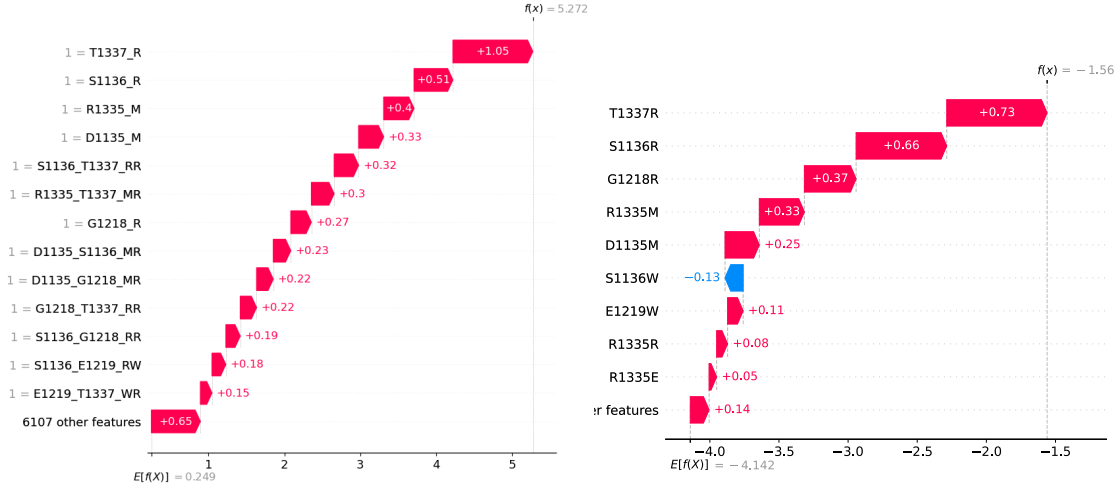
First, we trained a linear model with pairwise interaction features, and found that this did indeed improve model performance. The linear model is now nearly on par with the neural network; these results are described in the manuscript on page 9 and shown in **Sup. Fig. 9a** for the datasets labeled “one-hot with pairs”.

Second, we have added many new analyses related to the features of PAMmla enzymes that contributed to their modified PAM requirements. Our revised manuscript now includes SHAP analyses paired with structural modeling. We compared using classic SHAP in combination with the linear model as well as using DeepSHAP[21] combined with our neural network model. We preferred the output of DeepSHAP due to the smaller number of input features used with the neural network which we believe made the SHAP output more interpretable. These results are now included in our manuscript along with a substantial discussion of potential functional roles of amino acid substitutions in PAMmla enzymes (see page 13 in the manuscript and new **Sup. Note 6**, **Sup. Figs. 14** and **15a-l**). We also performed similar analyses for the specific enzymes utilized for *RHO* P23H targeting, MRRWMR and KRHWMR, identifying potential reasons for their specific preferences for NGTG over NGGG (now described on page 20 of our manuscript and **Figs. 6h-k**, **Sup. Figs. 30a-e,** and **Sup. Note 10**).



Supplementary Figure 9. Machine learning models to predict PAM profile from amino acid sequence. *(**a**) Comparison of machine learning model architectures (linear regression, random forest, and neural network) and amino acid encodings (one-hot, one-hot plus all pairwise amino acid combinations, and Georgiev[22]). The R^2 value is shown between the experimentally determined k (via HT-PAMDA) and the predicted k (via each ML model) for an internal 5-fold cross-validation on the training set. Each validation set is sub-divided according to the minimum hamming distance (HD) of each variant to the nearest neighbor in the corresponding training set; thus, validation sets become more challenging as HD increases.*

We compared SHAP analyses for the PAMmla-predicted MRRWMR variant acting on an NGTG PAM from the linear model to the neural network (**Rebuttal Fig. 1**). Because the neural network required a smaller feature set to generate accurate predictions (only single AA substitutions) whereas the linear model required all pairwise AAs as features, SHAP values were more readily interpretable for the neural network model. Predictions for the linear model were explained by small contributions from a large number of features for which it is difficult to extract meaning visually. Whereas predictions for the neural network model could be mainly accounted for by the top 5-10 features. We therefore decided to use our neural network model combined with DeepSHAP for the feature analyses performed in this study.



Rebuttal Fig. 1. SHAP values for the PAMmla-predicted MRRWMR variant acting on an NGTG PAM from the linear model (left) to the neural network (right).

I don't understand why the ISDE process is necessary? The authors have already computed the PAMmla predictions for all the 64 million sequences. Computing a score on these should be very fast and sorting 64 million numbers should take at most a few hours and possible much less using fast / parallel sorting algorithms, which would give a perfect knowledge of the landscape rather than requiring an optimization algorithm.

Response 3.4: Per the Referee's suggestion, since ISDE is not an essential point of novelty in our work, we have now moved much of the discussion of ISDE to a newly expanded **Sup. Note 8** in our revised manuscript. That being said, we respectfully advocate that ISDE is a useful tool for exploratory data analysis by the community, which we believe will make the dataset more accessible to a wider user base. Analysis of evolutionary trajectories and prioritization of optimal enzymes can be accomplished very rapidly using our PAMmla-ISDE web interface: https://pammla.streamlit.app/Evolve_Variants. ISDE allows users to almost instantaneously obtain enzymes with desired properties via an interactive tool for users from all computational backgrounds (in a way not possible when sorting large files), quickly identifying variant enzymes and refining search criteria with real-time feedback. Additionally, our website visualizes ISDE trajectories to displays enzymes along the entire evolution path to illustrate how various mutations interact and accumulate to result in the final PAM preference. We now demonstrate in new data that ISDE typically converges on the optimal variant (please see new **Sup. Figs. 23b-d** and ***Response 1.12***, above).

Although sorting the 64 million predictions may not be prohibitively slow, storing the predictions does require download or generation of the full prediction set which is currently a collection of csv files that require ~25 GB of storage. ISDE provides an accessible option for users who do not wish to download or manipulate such a dataset, and instead prefer a simple interface to navigate our library of PAMmla predictions.

Structural modeling of the variants obtained (at least the most interesting ones) would strengthen the manuscript.

Response 3.5: We thank the Referee for this comment! We have added many new analyses related to structural predictions of PAMmla enzymes to attempt to understand the determinants of their modified PAM interactions. Our revised manuscript includes SHAP analyses, structural modeling, and a substantial discussion of potential functional roles of amino acid substitutions in PAMmla enzymes (see page 13 in the manuscript and new **Sup. Note 6**, **Sup. Figs. 14** and **15a-l**). Furthermore, we performed similar analyses for specific enzymes described in our *RHO* P23 experiments (MRRWMR and KRHWMR) and their specific preferences for NGTG over NGGG. For these enzymes we performed SHAP analyses and structural modeling that are now described on page 20 of our manuscript and in **Figs. 6i-l**, **Sup. Figs. 30a-e,** and **Sup. Note 10**. Future studies may more thoroughly elucidate the functional roles of amino acid substitutions for more diverse classes of PAMmla enzymes.

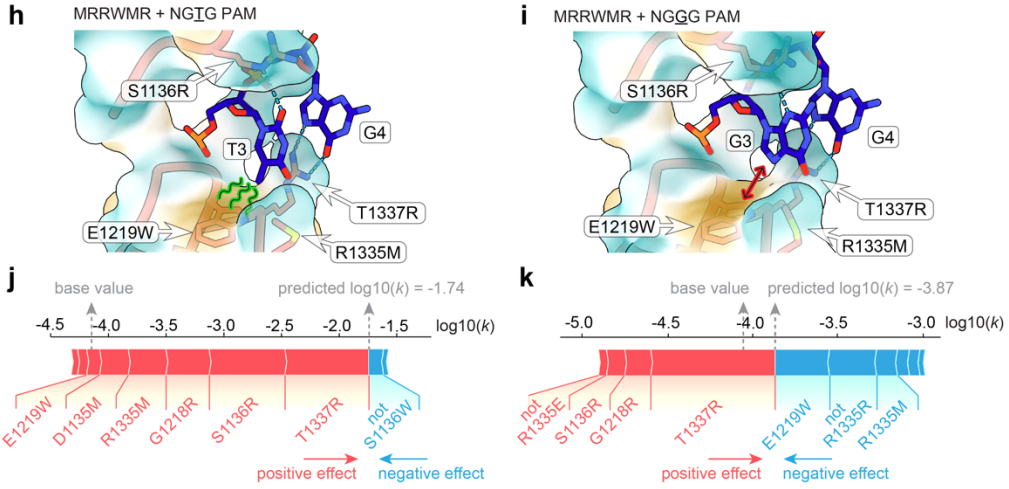


Figure 6. In silico directed evolution of an allele-specific editor for the RHO P23H allele. *(**h, i**) MRRWMR modelled on the structure of VRER (PDB: 5FW3)[8] interacting with NGTG or NGGG PAMs (**panels h** and **i**, respectively). Protein surface is colored by lipophilicity potential. Hydrogen bonds are represented by dashed lines and Van der Waals interactions are represented by green squiggles. (**j, k**) Force plots depicting SHapely Additive exPlanations[9] (SHAP) values for MRRWMR activity on NGTG or NGGG PAMs (**panels j** and **k**, respectively).*

Finally, one of the main take away message from the manuscript is that generating high quality data on mutant libraries can help model fitness landscapes and design improved protein variants. While the impact of the work is made higher because it was done on Cas9, generating possibly useful variants, this message in itself is not novel and similar principles have been applied to numerous proteins, including model proteins such as GFP (https://www.nature.com/articles/s41467-023-38099-z) and antibodies (e.g. https://www.nature.com/articles/s41467-023-39022-2 to take just one example).

Response 3.6: Although protein engineering and ML have been combined in other studies (many that we already cited in our original submission), we believe that a main point of novelty of our manuscript is the ability to generate custom PAM variant enzymes via protein engineering combined with ML. A key to the success of our study was our carefully curated training set of kinetic data for a large number of enzymes (~20-to-100-fold greater characterization of new enzymes compared to other typical Cas9 engineering manuscripts). That being said, the Referee's point is well-taken and we have revised our Discussion section to cite other previous studies (including the two manuscripts referred to above), and have also tried to better highlight what we believe are the main points of novelty of our study.

Minor comments:

The authors should discuss and cite related work including: https://doi.org/10.1371/journal.pcbi.1011621
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9039034/
https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.full.pdf

Response 3.7: We have cited these papers and added additional text to our Discussion section on pages 21 and 22. Results from these previous studies are consistent with analyses presented in our manuscript (formerly **Sup. Fig. 24**, now numbered **Sup. Fig. 31** in our resubmission). The related paragraph in our discussion section now reads as:

"We envision a continued necessity of high-quality experimental data to train ML models for certain protein engineering applications, like PAMmla. Alternatively, recent models trained on evolutionary sequence data have accurately predicted variant effects, sometimes with comparable performance to high throughput experimental assays[23,24]. However, we found that several evolutionary sequence-based models (Evcouplings[25], EVE[23], and TranceptEVE[24]) trained on 1,296 Cas9 related protein sequences from the UniRef100 sequence database were not predictive of enzyme activity on non-canonical

*PAMs (**Sup. Fig. 31**). Thus, there may be limitations of natural sequence-based approaches when evolving enzymes to perform non-natural functions, especially in cases where there is insufficient functional diversity or sequence homology amongst orthologs. In agreement with this result, previous ML-based approaches have sought to predict sequence-diversified but functional Cas9 enzymes[12,13] or to modify alternate Cas properties like the on-target activity of SaCas9-KKH[14,15] (**Sup. Note 12**). While these studies support the use of ML for Cas enzyme engineering, they also demonstrate that natural sequence data is likely insufficient to train models to predict SpCas9 enzyme variants with altered PAM requirements. The use of experimental data to train PAMmla overcomes this limitation by producing hundreds of non-natural SpCas9 proteins with properties not yet identified in nature."*

Initial screening in the bacterial work is restricted to NGNN PAMs. This choice should be better explained.

Response 3.8: We have added additional details related to our reasoning for modifying the 3rd and 4th positions of the PAM, including which amino acids we mutated (see also our ***Response 2.1***, above). Page 5 of our manuscript now reads:

*"Based on SpCas9 structures[7] and previous engineering efforts[1,2,10], we selected 6 amino acid residues within the SpCas9 PAM interacting (PI) domain to simultaneously randomize to all possible amino acids (**Fig. 1c**, **Sup. Fig. 2a**, and **Sup. Note 3**). We selected residues D1135, S1136, G1218, E1219, R1335, and T1337 for saturation mutagenesis (leading to an SpCas9(6AA) library of plasmids). These six amino acids modulate specificity for the 3rd and 4th PAM bases in the context of the engineered enzymes SpCas9-VRER, SpCas9-VRQR and SpG[1,2,10] (**Sup. Note 3**). Expansion of our library to also mutate R1333, the residue which contacts the second guanine of the SpCas9 PAM[11], largely abrogated enzyme activity (**Sup. Note 3**). We therefore focused our efforts on altering the 3rd and 4th positions of the PAM."*

The authors should discuss and cite related work by Malbranke et al., PLoS Comp. Biol. 2023 (https://doi.org/10.1371/journal.pcbi.1011621) as well as preprinted work by Ruffolo et al. (https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.full.pdf) if the journal allows.

Response 3.9: In our revised manuscript, we have included a brief discussion of these studies and their relation to our current work, and please see our ***Response 3.7***, above.

References

1. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* (2015) doi:10.1038/nature14592.
2. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science (1979)* (2020) doi:10.1126/science.aba8853.
3. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science (1979)* (2018) doi:10.1126/science.aas9129.
4. Miller, S. M. *et al.* Continuous evolution of SpCas9 variants compatible with non-G PAMs. *Nat Biotechnol* (2020) doi:10.1038/s41587-020-0412-8.
5. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* (2018) doi:10.1038/nature26155.
6. Kim, N. *et al.* Deep learning models to predict the editing efficiencies and outcomes of diverse base editors. *Nat Biotechnol* **42**, 484–497 (2024).
7. Kim, H. K. *et al.* High-throughput analysis of the activities of xCas9, SpCas9-NG and SpCas9 at matched and mismatched target sequences in human cells. *Nat Biomed Eng* **4**, 111–124 (2020).
8. Anders, C., Bargsten, K. & Jinek, M. Structural Plasticity of PAM Recognition by Engineered Variants of the RNA-Guided Endonuclease Cas9. *Mol Cell* (2016) doi:10.1016/j.molcel.2016.02.020.
9. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. doi:10.5555/3295222.3295230.
10. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 variants with undetectable genome-wide off-targets. *Nature* **529**, 490 (2016).
11. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* (2014) doi:10.1038/nature13579.
12. Malbranke, C. *et al.* Computational design of novel Cas9 PAM-interacting domains using evolution-based modelling and structural quality assessment. *PLoS Comput Biol* **19**, e1011621 (2023).
13. Ruffolo, J. A. *et al.* Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences. doi:10.1101/2024.04.22.590591.
14. Thean, D. G. L. *et al.* Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities. *Nature Communications 2022 13:1* **13**, 1–14 (2022).
15. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat Biotechnol* **33**, 1293–1298 (2015).
16. Ciciani, M. *et al.* Automated identification of sequence-tailored Cas9 proteins using massive metagenomic data. *Nature Communications 2022 13:1* **13**, 1–8 (2022).
17. Pedrazzoli, E. *et al.* CoCas9 is a compact nuclease from the human microbiome for efficient and precise genome editing. *Nature Communications 2024 15:1* **15**, 1–12 (2024).
18. Ma, E., Harrington, L. B., O'Connell, M. R., Zhou, K. & Doudna, J. A. Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Mol Cell* **60**, 398–407 (2015).
19. Chen, X., Liu, J., Janssen, J. M. & Gonçalves, M. A. F. V. The Chromatin Structure Differentially Impacts High-Specificity CRISPR-Cas9 Nuclease Strategies. *Mol Ther Nucleic Acids* **8**, 558 (2017).
20. Eggers, A. R. *et al.* Rapid DNA unwinding accelerates genome editing by engineered CRISPR-Cas9. *Cell* **187**, 3249-3261.e14 (2024).
21. Chen, H., Lundberg, S. M. & Lee, S. I. Explaining a series of models by propagating Shapley values. *Nature Communications 2022 13:1* **13**, 1–15 (2022).
22. Georgiev, A. G. Interpretable numerical descriptors of amino acid space. *Journal of Computational Biology* **16**, 703–723 (2009).
23. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature 2021 599:7883* **599**, 91–95 (2021).
24. Notin, P. *et al.* TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction. doi:10.1101/2022.12.07.519495.
25. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).

Response to Editorial Comments:

Your manuscript, "Design of custom CRISPR-Cas9 PAM variant enzymes via machine learning", has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are very interested in the possibility of publishing your study in Nature, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

We therefore invite you to revise your manuscript taking into account the following points. The most important aspects of the referee requests are to:

(1) release the datasets mentioned in the data availability statement (this must be done as a priority now, rather than at point of publication, so that referees can assess the data if needed);

Response: The primary datasets for our study are available in the Supplementary Tables, we have started the process of making them available on the NCBI SRA (and assure that all NGS results will be deposited there, this just takes some time), they are downloadable from GitHub, and the PAMmla predictions are available via a website. Accordingly, we have edited our Data Availability statement to read as follows to more explicitly elaborate on where users can access the datasets:

*"Primary datasets for this study are available in **Sup. Table 1** (HT-PAMDA data), **Sup. Table 2** (PAMmla predictions), and **Sup. Table 6** (GUIDE-seq2 data). The HT-PAMDA training data are also available on GitHub at https://github.com/RachelSilverstein/PAMmla. Next-generation sequencing results are available through the NCBI sequence read archive (SRA) under PRJNA1169103. PAMmla predictions for all 64 million SpCas9(6AA) enzymes can be viewed through an online webtool (https://pammla.streamlit.app/). Plasmids have been deposited with Addgene and are available at: https://www.addgene.org/Benjamin_Kleinstiver/ (see also **Sup. Table 4**)."*

(2) ensure that the Guide-Seq2 is preprinted as a priority (again, so that referees can examine the methodology, if needed);

Response: Per our discussion with the Editor, we have been communicating with Dr. Shengdar Tsai, and he he has assured us that they will preprint the GUIDE-seq2 manuscript this month (October). For utility to the community, our methods section succinctly describes how to perform GUIDE-seq2. Our manuscript does not claim to be the first to describe this method, preserving novelty for their eventual preprint and manuscript. Please also note that Dr. Tsai and the members of his lab involved in developing GUIDE-seq2 are co-authors on our manuscript.

Once the Tsai lab has preprinted their GUIDE-seq2 manuscript, we will cite it in our Methods section (where the text currently reads: "*GUIDE-seq2 reactions were performed essentially as described (Lazzarotto & Li et al, in preparation) with minor modifications.*", then followed by a sufficiently detailed description of the method).

(3) regarding generalized claims of lower off-target properties of the new Cas9 variants, this will either need extending beyond the 5 currently tested enzymes (ideally), or the text caveated/contextualized to acknowledge the limitations of the testing.

Response: In our revised manuscript, we had performed additional GUIDE-seq experiments, which together brought the total comparisons of PAMmla versus prior PAM-relaxed enzymes to ten separate comparisons [these datasets can be found in **Figs. 5c** (2 comparisons)**, 5d** (3 comparisons), **5g** (1 comparison), **6g** (2 comparisons), and **Extended Data Fig. 9c** (2 comparisons)]. We note that across all ten of these comparisons, the PAMmla enzymes have a reduced number of off-target sites and an improved ratio of on-to-off-target GUIDE-seq read counts compared to PAM-relaxed enzymes SpG and SpRY. The ten comparisons that we performed already

meets the threshold that the Referee requested, so we respectfully suggest that additional experiments are unlikely to meaningfully change our conclusions about these experiments.

Per our discussion with the Editor, we carefully reviewed our claims about specificity and generally feel that they were appropriate given our results. We do not claim or expect that all PAMmla enzymes reduce off targets since PAMmla is able to predict enzymes with a range of PAM preferences including those with relaxed preferences similar to SpG. We only claim that, for the narrowed PAM preference enzymes which we tested, we found reduced off-targets compared to SpG and SpRY. We have adjusted language to ensure that our claims about specificity improvements with PAMmla enzymes are balanced both in the main text and in Sup. Note 11.

(4) We also strongly suggest that your revised manuscript has tracked changes, which is increasingly requested by referees to aid in their re-review.

Response: We have tracked all changes to our manuscript.

Response to Referee #1

In this revised manuscript, Silverstein et al. have addressed and strengthened the proof for the key questions, including the engineering framework using ML and ISDE methods, and the structural analysis of the SpCas9 enzyme. The modified version is significantly improved and of high quality.

Response: We thank the Referee for their kind comments and the contributions to improving our work.

There is one additional question here that I would like to raise.

1. The number of enzymes analyzed for PAMmla-predicted enzymes was insufficient to demonstrate that these enzymes with altered PAMs can reduce genome-wide off-target effects compared to PAM-relaxed enzymes. Current results were based on five enzymes, which only show that some PAMmla enzymes with altered PAMs can reduce off-target effects. To fully assess the application of PAMmla on the specificity of the SpCas9 enzyme, a larger number of enzymes (at least 10-20, ideally over 50-100) would need to be analyzed. The authors may want to revise the related statements/claims to ensure validity.

Response: In our original revised manuscript, we had performed additional GUIDE-seq experiments as requested by Referees, which together brought the total comparisons of PAMmla versus prior PAM-relaxed enzymes to ten separate comparisons [these datasets can be found in **Figs. 5c** (2 comparisons)**, 5d** (3 comparisons), **5g** (1 comparison), **6g** (2 comparisons), and **Extended Data Fig. 9c** (2 comparisons)]. We note that across all ten of these comparisons, the PAMmla enzymes have a reduced number of off-target sites and an improved ratio of on-to-off-target GUIDE-seq read counts compared to PAM-relaxed enzymes SpG and SpRY. The ten comparisons already meet the requested threshold, so we respectfully suggest that additional experiments are unlikely to meaningfully change our conclusions about these experiments and enzymes.

Per our discussion with the Editor, we carefully reviewed our claims about specificity and generally feel that they were appropriate given our results. We feel that we haven't made a major emphasis on this point in our manuscript to ensure, but we have also adjusted language pertaining to our claims about specificity improvements with PAMmla enzymes to be more balanced both in the main text and in Sup. Note 11.

The PAM activity prediction website (https://pammla.streamlit.app) was working but is sleeping due to inactivity

Response: The website works hard and sometimes needs a rest! Please note that the website can be easily 'awoken' by clicking on the button that will wake up the website (which will active all functions within ~1 minute).

Response to Referee #2

The authors submitted a revised version of their manuscript that adequately addressed my previous comments. They discuss effects of varying the 2nd PAM position and added an analysis to investigate the importance of oversampling. I enjoyed the added sections on structure modelling approaches to understand the molecular basis for altered PAM requirement. Finally, two references for very recent papers with alternative ML approaches to generate Cas9 variants with modified PAM specificity have been added. The citation of the unpublished GUIDE-seq2 work and the final data availability statement will need to be discussed with the Nature editorial board.

Response: We are gratified to learn that the Referee finds our revisions appropriate and thank them for their helpful comments and effort in guiding our revisions.

Regarding GUIDE-seq2, we discussed this with the Editor; once the Tsai lab has preprinted their GUIDE-seq2 manuscript, we will cite it in our Methods section (where the text currently reads: "*GUIDE-seq2 reactions were performed essentially as described (Lazzarotto & Li et al, in preparation) with minor modifications.*", then followed by a sufficiently detailed description of the method).

The primary datasets for our study are available in the Supplementary Tables, we have started the process of making them available on the NCBI SRA (and assure that all NGS results will be deposited there, this just takes some time), they are downloadable from GitHub, and the PAMmla predictions are available via a website. Accordingly, we have edited our Data Availability statement to read as follows to more explicitly elaborate on where users can access the datasets: *"Primary datasets for this study are available in **Sup. Table 1** (HT-PAMDA data), **Sup. Table 2** (PAMmla predictions), and **Sup. Table 6** (GUIDE-seq2 data). The HT-PAMDA training data are also available on GitHub at https://github.com/RachelSilverstein/PAMmla. Next-generation sequencing results are available through the NCBI sequence read archive (SRA) under PRJNA1169103. PAMmla predictions for all 64 million SpCas9(6AA) enzymes can be viewed through an online webtool (https://pammla.streamlit.app/). Plasmids have been deposited with Addgene and are available at: https://www.addgene.org/Benjamin_Kleinstiver/ (see also **Sup. Table 4**)."*

Response to Referee #3

The authors have satisfactorily addressed my all comments. The manuscript is now much better than the initial submission. I support publication and congratulate the authors on a very nice piece of work. My only remaining concern might be with the title which in my opinion places too much emphasis on machine learning, which is not particularly original in this study, and not enough on the dataset which is the key contribution.

Response: We are grateful to the Referee for their helpful comments during the initial review and are glad to hear that our responses were satisfactory.

Regarding the title, thanks for this comment. ML is an integral aspect of our study and was key to developing the resulting variant enzymes. We agree that our large-scale dataset is also a key aspect of our manuscript, so we have modified the title to: "Design of custom CRISPR-Cas9 PAM variant enzymes via scalable engineering and machine learning". We feel that this new title appropriately reflects the dual importance of both experimental and ML aspects of our work. However, this title is now over the character limit, so we will need to discuss the use of this new title with the editorial team.

I didn't try to run the code myself, but it seems to be well documented and a useful ressource for the community.

Response: Thank you!