nature portfolio

Peer Review File

Training Data Diversity Enhances the Basecalling of Novel RNA Modification-Induced Nanopore Sequencing Readouts

Corresponding Author: Dr Hongxu Ding

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors tried to retrain ONT Direct RNA sequencing reads to minimize the errors caused by modifications. They found that trained models had a better generalization when enough modification types were introduced. The findings may indicate a better way of training RNA basecall models for ONT reads. However, the findings are empirical, and the reasoning behind them is not well discussed. Also, due to the fast development of ONT related tools, the authors need to provide more convincing evidence that their approach could be applied to other datasets.

Major points:
1. Why ac4C, Psi and m1Psi bring the more basecalling errors in the analysis? Do these three modifications pose larger signal deviations from canonical bases? I suggest the authors showcase the current feature (dwell time, mean/median and SD) of these modification sites (Tombo plot function or NanoCEM), and discuss the possible reasoning in the discussion.

2. ONT has replaced the LSTM in their basecallers (Guppy6 or Dorado 0.6.x) to transformer (Dorado 0.7). Can the authors showcase (or maybe discuss) that if their approach can still be useful in the new system?

3. The curlcake dataset is good for initiating a test. However, can the author showcase if their retrained model performed better in any real datasets? The authors may use tRNAs or rRNAs, which are abundant in modifications, as a real test dataset.

4. Figure 3 & Result 5 : if using only bad representation(m5C& m5U & Psi & m1Psi), can the authors still generalize the representation space to out-of-sample modifications? How about the observed accuracy of UM after being trained by ALL? Will it decrease?

Minor points:

1. Figure1 (and all figures), please add detailed labels for the meaning of words used in the Figure in the legends. Like: “Novel” indicates novel modifications not used in the training, or novel artificial nucleotides, or novel kmers for canonical nucleotides? What is the meaning of “++”?

2. Figure 2A, not sure why the x-axis for deletion could be as high as 25%, but insertion and mismatch are 5-6%. Does the author have any reasoning behind this? The deletion ratio in Nanopore direct-RNA is 2-3% higher than insertion, but 5 times higher may indicate some over estimation in counting methods.
Figure 2C The Label “Self” means this specific type of Modification data is used for training, and “All” means all modifications are used. Then I would suggest using AllMod, onemod or Unmod.
Figure 2B, Add the color coding.
Also figure 2C, I would also suggest the author provide three distributions separately. The bidirectional density plot for three samples is confusing.
Figure 2D, I can not read information directly from the figure, does the author indicate the positions are aligned or unaligned?

Why is the UM missing from the positional analysis? The author also needs to present the difference in the number of bases for the same read using different basecaller.

(Remarks on code availability)
It is hard to run the code nowadays because the installation of some dependencies, such as taiyaki has been marked as "unsupported" from ONT and requires old versions of Python. The authors may need to provide docker for users to reproduce their work.

Reviewer #2

(Remarks to the Author)
In this manuscript, the authors focused on the basecalling of RNA oligos in the context of modifications. They demonstrate that the performance of the basecalling model can be improved by incorporating diverse training data containing different modifications. They attribute the enhancement to a better representation space of the latent vectors generated by the encoder.

Here are the comments for the current manuscript:

- The authors trained their own models with Guppy and made comparisons. How about the performance of the latest basecallers (e.g., Dorado) on the same test set? Although it is difficult to make strict comparisons, demonstrating the performance of commonly used basecallers can be useful as a reference.

- The authors randomly sampled approximately 20,000 reads for each modification type and combined them together, resulting in balanced basecall datasets with and without modifications. Since the diversity of training data is key for this work, and in real scenarios, data imbalance usually exists. Is data balancing a prerequisite for the proposed method? If not, how does imbalanced data with diversity affect the method's performance?

- In the manuscript, "We found that ac4C and training reads tend to co-localize in “UM”, “m1A”, “m6A” and “hm5C” representation space, as opposed to the error-prone “m5C”, “m5U”, “Psi” and “m1Psi” (Figure 3B)."
How is "co-localization" assessed, especially for "m1A" and "m6A"?

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The article by Ziyuan et al. discussed about the basecalling accuracy when dealing with the nanopore sequencing reads in presence of the nucleotide modifications. The authors tested the results of the Guppy basecaller by training with different types of data, including the unmodified RNA and data with diverse modifications. The main conclusion is that exposure of the basecaller to diverse data improves the accuracy of basecalling. The study is a nice technical test, and the manuscript can be a good technical note. However, the manuscript does not have the methodology scope, the novelty of a new method, or an in-depth analysis for a biological question to warrant its publication. It is still preliminary and needs quite a lot of work to make a more complete methodology or analysis article.

1. The authors claimed that “increasing the training data diversity as a novel paradigm for building modification-tolerant nanopore sequencing basecallers”, which is not precise. Similar notes have been made before, e.g., https://www.biorxiv.org/content/10.1101/2023.11.28.568965v1. The ONT company is also providing basecallers trained with native RNA, which includes diverse modifications. Therefore, it is not novel at all to simply train the models with diverse data.
2. The technical tests are too simple. The authors simply compared the results of basecalling upon training with unmodified vs. modified RNA data. No comparison has been done with other pipelines of basecalling.
3. Analyses of the representation space are also descriptive, and the results are not surprising at all. The authors did not bother to discuss about why adding some modifications improved the basecalling accuracy but adding some other types did not.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors' revision has significantly improved the quality of the manuscript. However, some issues remain, particularly regarding the explanation of the method's generalization and the accommodation of new chemistries and models.

1. The authors retrained the basecalling model using the tRNA dataset to demonstrate improved tRNA basecalling accuracy. However, this improvement might result from learning specific k-mers within a small dataset. Moreover, this approach does not address the question: Can a model retrained from the curlcake dataset improve tRNA basecalling accuracy? The generalization capability of the trained model needs to be demonstrated through different approaches.

2. Much of the added discussion is somewhat shallow and computationally focused, lacking sufficient consideration of the underlying reasoning behind the findings. Please consider rewriting these sections to provide deeper insights. For example, in response to the observation that "deletion rates could be as high as 25%, while insertion and mismatch rates are 5–6%," the authors stated "high-level signal deviations." However, significant signal deviations would likely cause more mismatches along with indels, resulting in a proportional increase in both mismatches and indels. High indel rates may be related to changes in dwell time, but a higher rate of deletions rather than insertions could also be algorithm-related. There are many interesting details within these findings, but the authors have overlooked other possibilities without providing sufficient reasoning.

Minor Point:

Consider replacing Figure 2D with two or more one-dimensional distributions (e.g., violin plots with boxplots) representing the mapped length ratio (0–100%). In the current figure, I can only discern the highlighted differences between X and Y.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed the technical concerns.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I am afraid that my previous concerns have not been fully resolved. In the revised manuscript, the authors have added some more analyses with other pipelines. However, my concern lies in the methodology design and originality.

The major point of this study is that training of data with diverse modifications generalize the representation space, so that the basecalling for the nts with out-of-sample modifications would be more accurate. As I have mentioned in my previous comments, this idea itself is not original. Similar strategy has been used in other studies (e.g., the preprint on biorxiv, which presents a more novel method for detections of RNA modifications). Technical notes on ONT have also recommended such a strategy. I disagree with the authors on their claim that such training model has yet to be investigated. It is not "a novel paradigm for training modification-agonistic basecallers".

The authors used different data to demonstrate that using diverse raw data for training can improve basecalling accuracy. This is empirical. Again, this work would be a nice technical note, but not a method paper with novel design or new concepts of methodology design. I simply don't think it meets NC's standard for innovation.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed my concerns.

(Remarks on code availability)

Reviewers' comments:

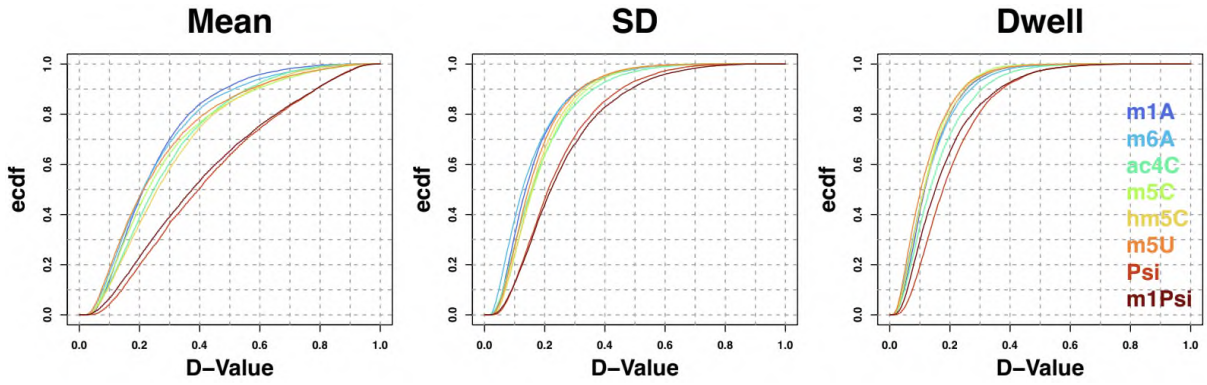
Reviewer #1 (Remarks to the Author):

The authors tried to retrain ONT Direct RNA sequencing reads to minimize the errors caused by modifications. They found that trained models had a better generalization when enough modification types were introduced. The findings may indicate a better way of training RNA basecall models for ONT reads. However, the findings are empirical, and the reasoning behind them is not well discussed. Also, due to the fast development of ONT related tools, the authors need to provide more convincing evidence that their approach could be applied to other datasets.

Major points:

1. Why ac4C, Psi and m1Psi bring the more basecalling errors in the analysis? Do these three modifications pose larger signal deviations from canonical bases? I suggest the authors showcase the current feature (dwell time, mean/median and SD) of these modification sites (Tombo plot function or NanoCEM), and discuss the possible reasoning in the discussion.

The reviewer is correct that ac4C, Psi and m1Psi modifications “pose larger signal deviations from canonical bases”, therefore could “bring the more basecalling errors in the analysis”. As the reviewer suggested, we used Remora (the state-of-the-art algorithm for signal-sequence alignment) to extract per-base current mean, SD and dwell time. We determined distributions of these features in different modifications, next quantified their deviations against canonical bases using KS-test D-values. As shown in the following figure, Psi and m1Psi could largely deviate all the three current features, while ac4C majorly affects the dwell time.



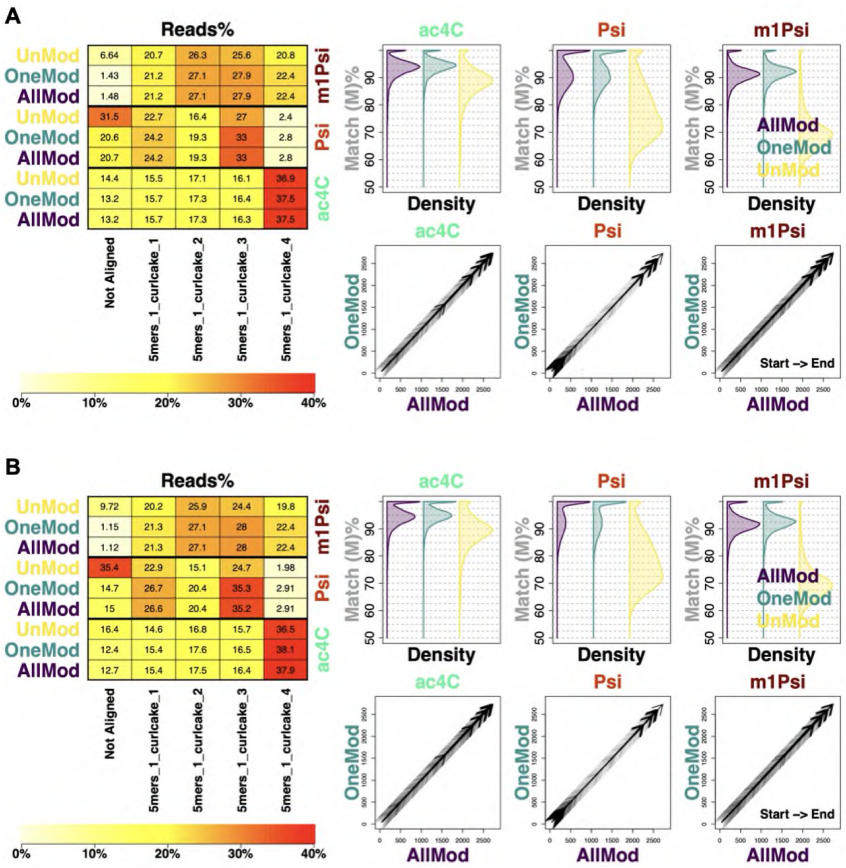
We included these results in the updated Figure S2, and updated the DISCUSSION section:

“*Promoting the basecalling accuracy, especially for native DNA/RNA sequencing signals, remains a central challenge in nanopore sequencing bioinformatics. This is because the latest basecallers are less tolerant to modifications, which commonly exist among native DNA and RNA molecules and will deviate nanopore sequencing signals (Figure S2). To properly address this limitation, basecallers that are agnostic to nucleotide modifications are therefore in urgent need.*”

2. ONT has replaced the LSTM in their basecallers (Guppy6 or Dorado 0.6.x) to transformer (Dorado 0.7). Can the authors showcase (or maybe discuss) that if their approach can still be useful in the new system?

With all respect, we would like to explain that the comparison with transformer-based models is less feasible, because these models were only compatible with the latest RNA chemistry (corresponding flowcells and kits are still early access products), while datasets analyzed in our study were sequenced using the R9.4.1 chemistry.

Nevertheless, we agree entirely with the reviewer that “if their approach can still be useful in the new system” needs to be demonstrated. We therefore systematically investigated Bonito and Dorado basecalling systems (panel A and B in the following figure, respectively), using configurations that are compatible with the R9.4.1 chemistry. We used mappability, per-read CIGAR match fraction and alignment consistency as quantification metrics, as in Figure 2B, C and D respectively. We noticed that basecallers trained with diverse modifications (AllMod) out-performed “negative control” basecallers trained with only unmodified data (UnMod), and achieved similar performance with “positive control” basecallers trained with the modification to be basecalled (OneMod). We thus confirmed our central conclusion that “training data diversity enhances the basecalling of novel RNA modification-induced nanopore sequencing readouts” with the latest basecallers. We included these results in the updated Figure S1.



3. The curlcake dataset is good for initiating a test. However, can the author showcase if their retrained model performed better in any real datasets? The authors may use tRNAs or rRNAs, which are abundant in modifications, as a real test dataset.

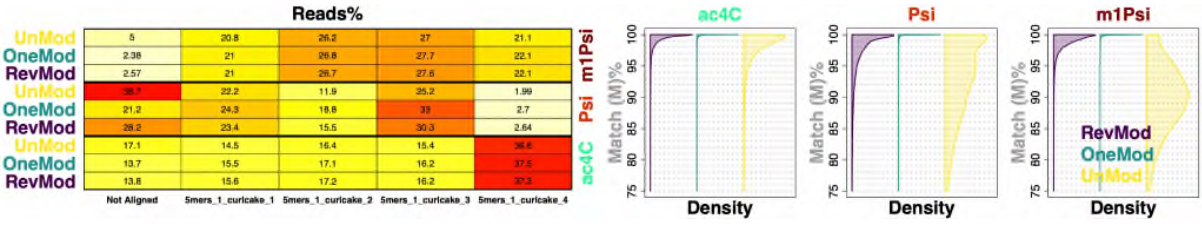
Following the reviewer's comment, we analyzed "tRNAs as a real test dataset". To prove that training data diversity enables modification-tolerant basecalling, without losing generality, we analyzed the most densely-modified yeast tRNA species Leu-TAA. We trained 1) a "positive" model with all the yeast tRNA species (except for Leu-TAA); 2) a "negative" model using only the 6 non-modified yeast tRNA species (Pro-AGG, Leu-AGG, Gly-CCC, Gln-CTG, Gln-TTG, Ala-TGC) to basecall Leu-TAA reads (9,918 in total). Nanopore sequencing data of wild-type yeast native tRNAs, and corresponding modification annotations, were downloaded from the Nano-tRNAseq paper (https://www.nature.com/articles/s41587-023-01743-6).

As in Figure 2, we quantified basecalling performance with mappability (number of reads that can be aligned to the reference sequence). The rationale is that excessive basecalling errors might disrupt alignment completely, further resulting in "unmappable" reads. As shown in the following table, the positive model which was trained by diverse modifications, accomplished a ~15% increase in mappability. We therefore concluded that our "retrained model performed better in real datasets".

	Not Aligned	Leu-TAA
Positive Model	716 (7.22%)	9,202 (92.78%)
Negative Model	2,154 (21.72%)	7,764 (78.28%)

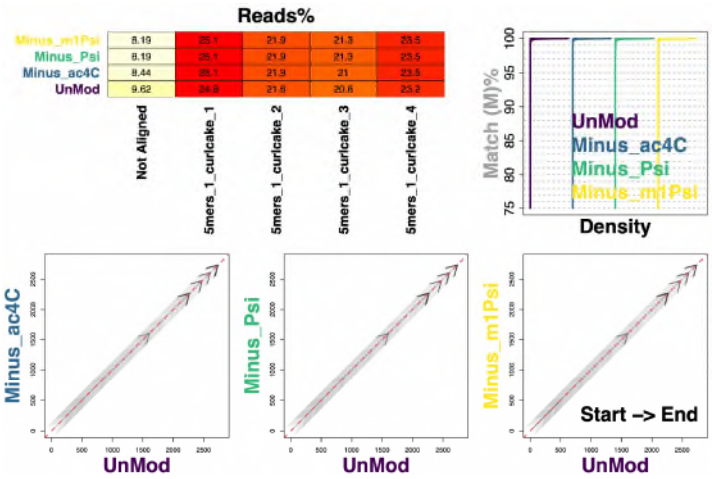
4. Figure 3 & Result 5 : if using only bad representation(m5C& m5U & Psi & m1Psi), can the authors still generalize the representation space to out-of-sample modifications? How about the observed accuracy of UM after being trained by ALL? Will it decrease?

Following the reviewer's comment, we explored the out-of-sample basecalling performance of "bad" modification combinations (RevMod). As the reviewer mentioned, for ac4C analysis, we considered {m5C, m5U, Psi, m1Psi} as the RevMod. We further analyzed Psi and m1Psi, and considered {UM, m6A, ac4C, m5C} to be the RevMod for both groups. As shown in the following figure, RevMod slightly decreased mappability, and largely decreased basecalling accuracy. We therefore concluded the compromised out-of-sample generalizability for "bad" training modification combinations.



We further examined whether "accuracy of UM after being trained by ALL" will decrease. As shown in the following figure, Minus_ac4C, Minus_Psi and Minus_m1Psi represent the three "ALL" basecallers, while UnMod represents the basecaller trained with only unmodified data,

as the baseline control. We ran the four basecallers on unmodified data, and found marginal mappability and basecalling accuracy differences. We further observed consistent alignment between ALL and UnMod groups. We therefore concluded that the ALL-training introduces no obvious decrease for the UM-accuracy.



Minor points:

1. Figure1 (and all figures), please add detailed labels for the meaning of words used in the Figure in the legends. Like: “Novel” indicates novel modifications not used in the training, or novel artificial nucleotides, or novel kmers for canonical nucleotides? What is the meaning of “++”?

We apologize for the confusion, and updated the Figure 1 legend for clarification:

“*UnMod and Mod denote unmodified and modified training data categories, respectively. Novel denotes the out-of-sample modification in the test data. Encoder++ and Decoder++ comprise the basecaller trained with diverse modifications, as opposed to the basecaller trained with only the UnMod data.*”

2. Figure 2A, not sure why the x-axis for deletion could be as high as 25%, but insertion and mismatch are 5-6%. Does the author have any reasoning behind this? The deletion ratio in Nanopore direct-RNA is 2-3% higher than insertion, but 5 times higher may indicate some over estimation in counting methods.

We thank the reviewer for the comment, and would like to explain that the high deletion rate, especially in the ac4C, Psi and m1Psi oligos, are most likely caused by significantly deviated nanopore sequencing signals. As the reviewer mentioned in Comment #1, which was proved by our analyses in the corresponding response, that ac4C, Psi and m1Psi “pose larger signal deviations from canonical bases”. Such high-level signal deviations could drastically bias the basecaller trained with only unmodified data, further introducing excessive deletions.

Figure 2C The Label “Self” means this specific type of Modification data is used for training, and “All” means all modifications are used. Then I would suggest using AllMod, onemod or Unmod.

We thank the reviewer for the comment, and updated Figure 2C together with other related figures accordingly.

Figure 2B, Add the color coding.

We thank the reviewer for the comment, and added the color coding in Figure 2B together with other related figures.

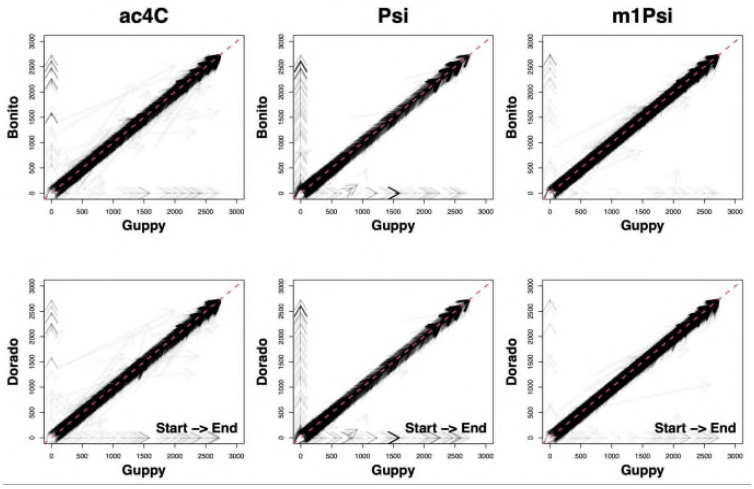
Also figure 2C, I would also suggest the author provide three distributions separately. The bidirectional density plot for three samples is confusing.

We thank the reviewer for the comment, and presented different distributions separately in Figure 2C together with other related figures.

Figure 2D, I can not read information directly from the figure, does the author indicate the positions are aligned or unaligned? Why is the UM missing from the positional analysis? The author also needs to present the difference in the number of bases for the same read using different basecaller.

We thank the reviewer for the comment, and would like to explain that Figure 2D shows the alignment consistency. Specifically, for each **aligned** read, an arrow from its alignment start position to end position was drawn (unaligned reads were ignored because they don’t have such positions). We considered basecallers trained with the modification to be basecalled (OneMod) as positive controls, thus compared basecallers trained with diverse modifications (AllMod) against them. Basecallers trained with only unmodified data (UnMod), on the other hand, are considered as less accurate negative controls, thus excluded from comparisons.

Following the reviewer’s comment, we next assessed alignment consistency among Guppy, Bonito and Dorado. Specifically, we performed the comparison on AllMod groups of ac4C, Psi and m1Psi, and observed highly-consistent alignments, as shown in the following figure:



Reviewer #1 (Remarks on code availability):

It is hard to run the code nowadays because the installation of some dependencies, such as taiyaki has been marked as "unsupported" from ONT and requires old versions of Python. The authors may need to provide docker for users to reproduce their work.

Following the reviewer's comment, we uploaded the Guppy-Taiyaki singularity container onto figshare, and provided the link in our GitHub (https://github.com/wangziyuan66/NanoRL).

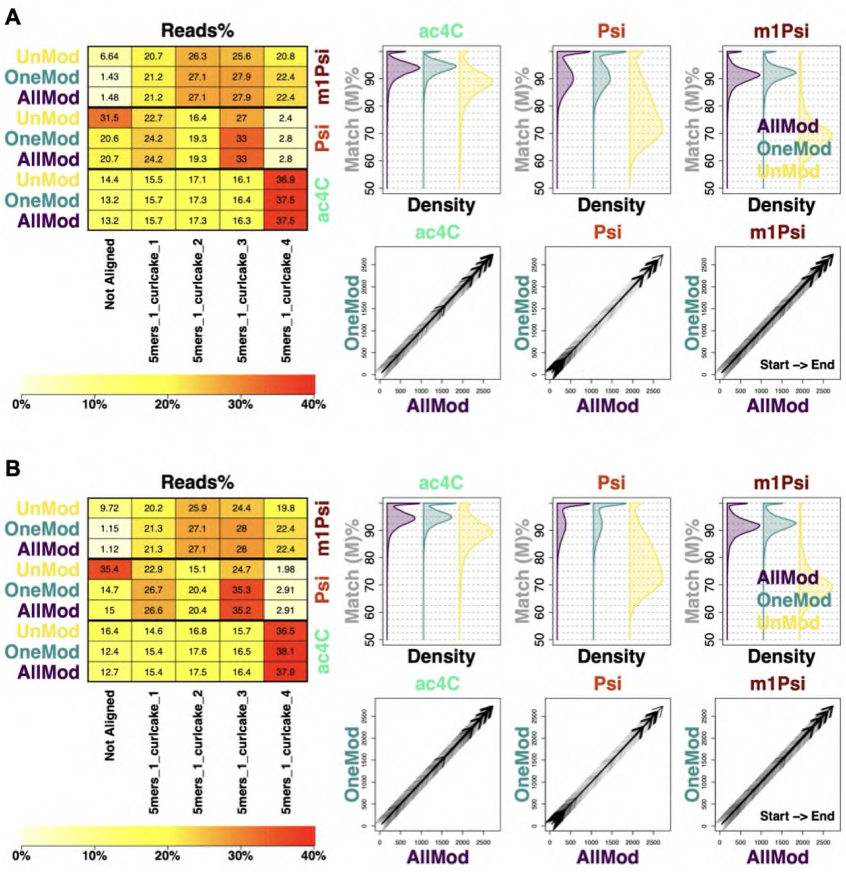
Reviewer #2 (Remarks to the Author):

In this manuscript, the authors focused on the basecalling of RNA oligos in the context of modifications. They demonstrate that the performance of the basecalling model can be improved by incorporating diverse training data containing different modifications. They attribute the enhancement to a better representation space of the latent vectors generated by the encoder.

Here are the comments for the current manuscript:

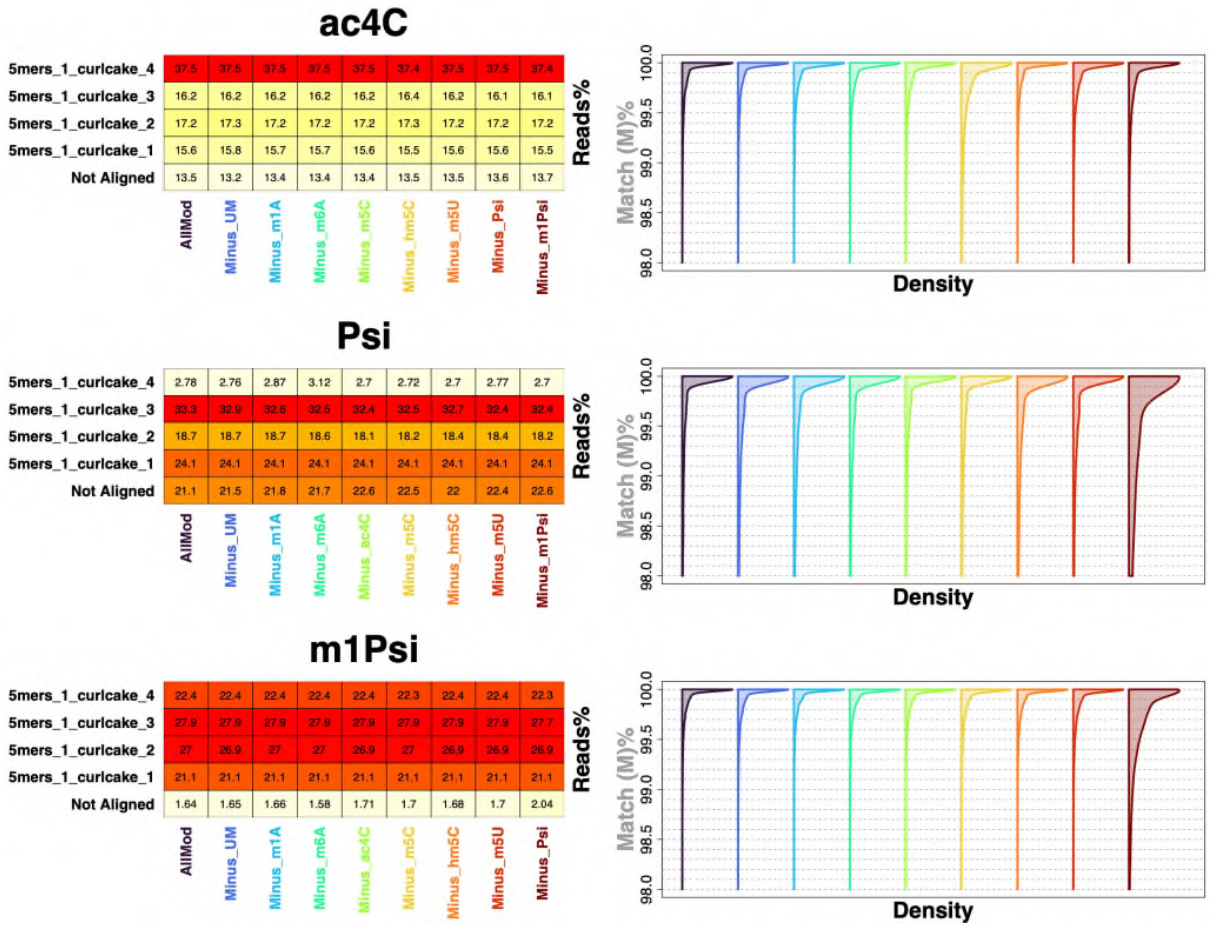
- The authors trained their own models with Guppy and made comparisons. How about the performance of the latest basecallers (e.g., Dorado) on the same test set? Although it is difficult to make strict comparisons, demonstrating the performance of commonly used basecallers can be useful as a reference.

We thank the reviewer for the comment, and systematically investigated the latest Bonito (A) and Dorado (B) basecallers, as shown in the following figure. We used mappability, per-read CIGAR match fraction and alignment consistency as quantification metrics, as in Figure 2B, C and D respectively. We noticed that basecallers trained with diverse modifications (AllMod) out-performed “negative control” basecallers trained with only unmodified data (UnMod), and achieved similar performance with “positive control” basecallers trained with the modification to be basecalled (OneMod). We thus confirmed our central conclusion that “training data diversity enhances the basecalling of novel RNA modification-induced nanopore sequencing readouts” with the latest basecallers. We included these results in the updated Figure S1.



- The authors randomly sampled approximately 20,000 reads for each modification type and combined them together, resulting in balanced basecall datasets with and without modifications. Since the diversity of training data is key for this work, and in real scenarios, data imbalance usually exists. Is data balancing a prerequisite for the proposed method? If not, how does imbalanced data with diversity affect the method's performance?

The reviewer is correct, that “data balancing a prerequisite for the proposed method”. To demonstrate such a conclusion, we investigated a group of extreme cases, where a certain modification type is completely removed from training. We denoted these cases as “Minus”. For example, “Minus_UM” stands for the removal of unmodified reads from training data. We investigated all the “Minus” groups for ac4C, Psi and m1Psi test reads, and summarized the results in the following figure:

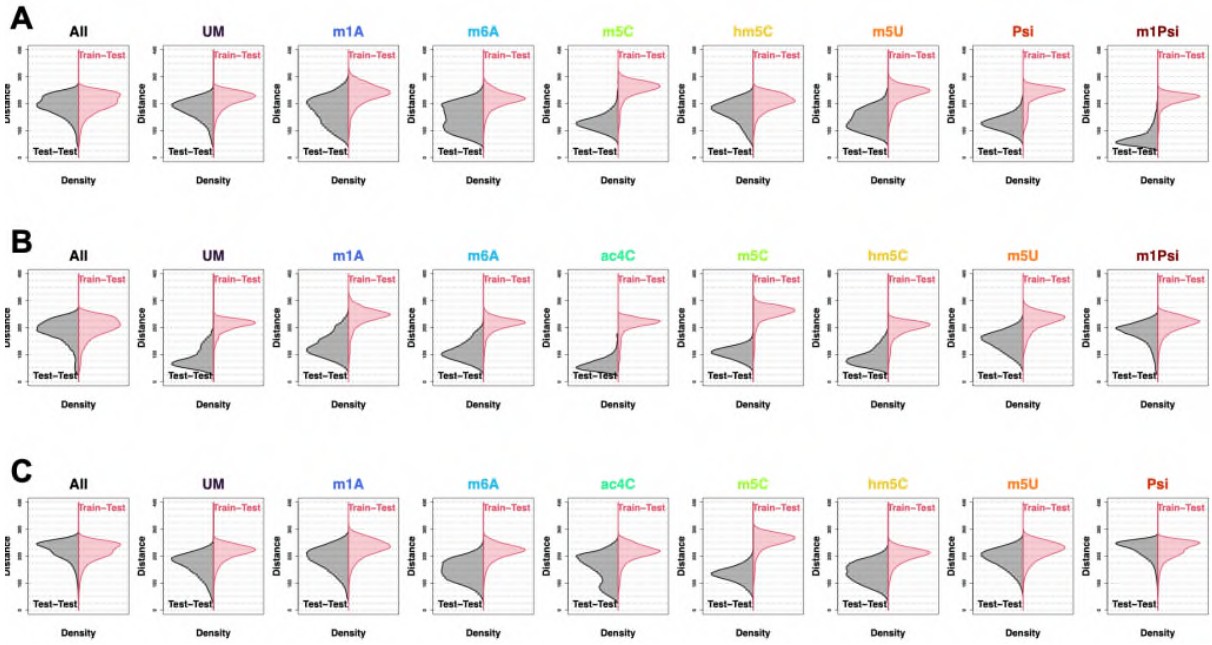


While the single modification-removal generally produced negligible effects on out-of-sample basecalling, we noticed that certain groups, e.g. Minus_m1Psi in Psi analysis and Minus_Psi in m1Psi analysis, yielded more pronounced basecalling performance decrease compared to their counterparts. We therefore concluded that imbalance training data, especially when key modifications are under-represented, might compromise out-of-sample basecalling.

- In the manuscript, "We found that ac4C and training reads tend to co-localize in “UM”, “m1A”, “m6A” and “hm5C” representation space, as opposed to the error-prone “m5C”, “m5U”, “Psi” and “m1Psi” (Figure 3B)."

How is "co-localization" assessed, especially for "m1A" and "m6A"?

We thank the reviewer for the comment, and acknowledge that quantitative assessments are required to fully support our conclusion that “precise basecalling requires high-quality data representations”. The UMAP co-localization, on the other hand, can only provide descriptive evidence. We thus, in the representation space, measured the inter-read Euclidean distance, to quantify the read-level similarity. We presented results for ac4C, Psi and m1Psi test reads in panel A, B and C of the following figure, respectively:



In the figure, All represents the jointly-trained basecaller by all the oligo groups except for the test, and other acronyms represent individually-trained basecallers. Train-Test and Test-Test denote the distance between training and test reads and among test reads, respectively. We took Test-Test groups as baselines to quantify Train-Test similarities: **closer Train-Test and Test-Test distributions indicate Test reads can be properly represented by Train reads**.

With the ac4C analysis (panel A) as an example, we noticed that in the more accurate “UM”, “m1A”, “m6A” and “hm5C” basecallers, test reads were more similar to training reads. On the contrary, test-train similarity was significantly reduced within error-prone “m5C”, “m5U”, “Psi” and “m1Psi” basecallers. Most importantly, the highest test-train similarity was produced by the most precise “All” basecaller. We therefore concluded that a high-quality representation space, which is determined by the high test-train similarity, is the prerequisite for accurate basecalling. We included such results in the updated Figure S3, and revised the RESULTS section “Precise basecalling requires high-quality data representations” accordingly:

“*Within the representation space, we quantified the read-level similarity and found that, in the more accurate “UM”, “m1A”, “m6A” and “hm5C” basecallers, ac4C test reads were more similar to their corresponding training reads. On the contrary, test-train similarity was significantly reduced among error-prone “m5C”, “m5U”, “Psi” and “m1Psi” basecallers. Most importantly, the highest test-train similarity was produced by the most precise “All” basecaller (Figure S3A). We further visualized such a representation space similarity pattern between*

test and training reads with UMAP, as shown in Figure 4B. In representation learning, test data points can be properly encoded, if and only if they fall in the manifold produced by training data points. We therefore highlight the importance of a high-quality representation space in precise basecalling. We further confirmed this conclusion with Psi (Figure S3B, S4A-C) and m1Psi (Figure S3C, S5A-C) test oligos.”

Reviewer #3 (Remarks to the Author):

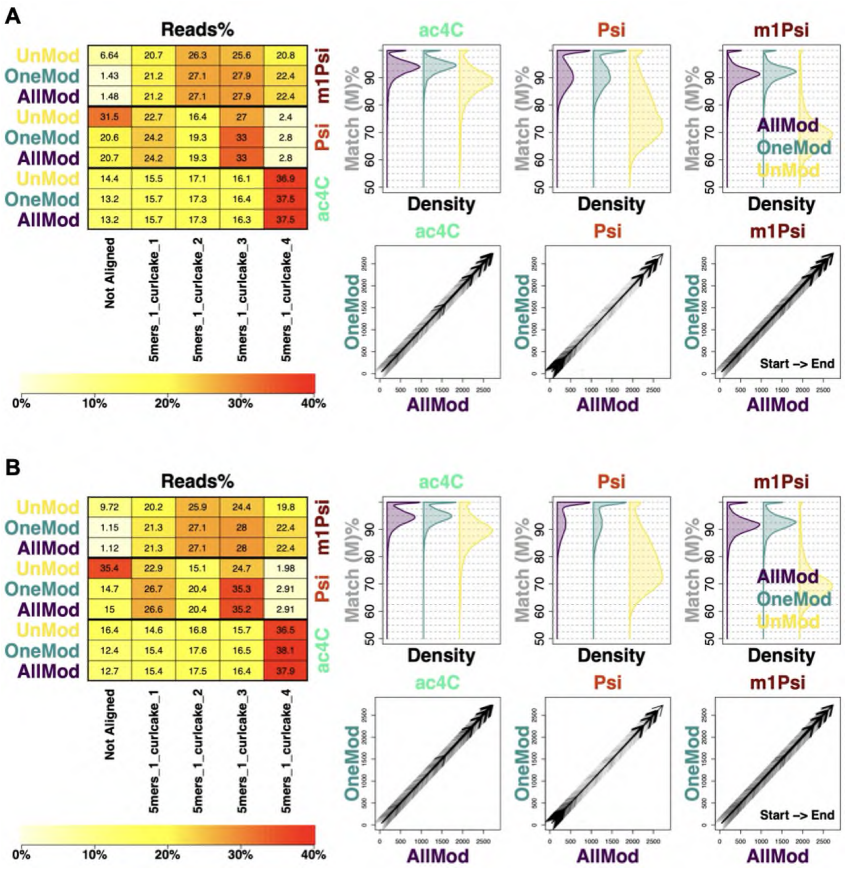
The article by Ziyuan et al. discussed about the basecalling accuracy when dealing with the nanopore sequencing reads in presence of the nucleotide modifications. The authors tested the results of the Guppy basecaller by training with different types of data, including the unmodified RNA and data with diverse modifications. The main conclusion is that exposure of the basecaller to diverse data improves the accuracy of basecalling. The study is a nice technical test, and the manuscript can be a good technical note. However, the manuscript does not have the methodology scope, the novelty of a new method, or an in-depth analysis for a biological question to warrant its publication. It is still preliminary and needs quite a lot of work to make a more complete methodology or analysis article.

1. The authors claimed that “increasing the training data diversity as a novel paradigm for building modification-tolerant nanopore sequencing basecallers”, which is not precise. Similar notes have been made before, e.g., https://www.biorxiv.org/content/10.1101/2023.11.28.568965v1.<https://www.biorxiv.org/content/10.1101/2023.11.28.568965v1.> The ONT company is also providing basecallers trained with native RNA, which includes diverse modifications. Therefore, it is not novel at all to simply train the models with diverse data.

We agree with the reviewer entirely that certain latest basecalling models were trained using diverse modifications. However, we would like to clarify that whether such models are able to properly handle **out-of-sample modifications** has yet to be rigorously investigated. Indeed, such a capability is essential for “modification-agnostic basecalling”, because we, in general, have no prior knowledge on novel modifications in real-world usecases. Precise basecalling, on the other hand, is the prerequisite for virtually all the downstream bioinformatic analyses. Our research closes such a knowledge gap, and demonstrates a novel paradigm for training modification-agonistic basecallers. We therefore respectfully disagree with the reviewer that our research is “not novel at all”.

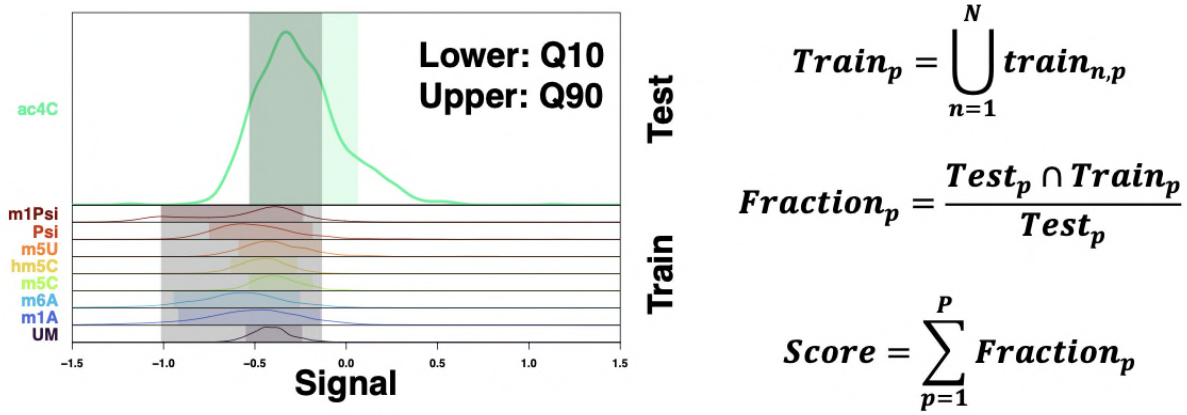
2. The technical tests are too simple. The authors simply compared the results of basecalling upon training with unmodified vs. modified RNA data. No comparison has been done with other pipelines of basecalling.

We thank the reviewer for the comment, and systematically investigated the latest Bonito (A) and Dorado (B) basecallers, as shown in the following figure. We used mappability, per-read CIGAR match fraction and alignment consistency as quantification metrics, as in Figure 2B, C and D respectively. We noticed that basecallers trained with diverse modifications (AllMod) out-performed “negative control” basecallers trained with only unmodified data (UnMod), and achieved similar performance with “positive control” basecallers trained with the modification to be basecalled (OneMod). We thus confirmed our central conclusion that “training data diversity enhances the basecalling of novel RNA modification-induced nanopore sequencing readouts” with the latest basecallers. We included these results in the updated Figure S1.

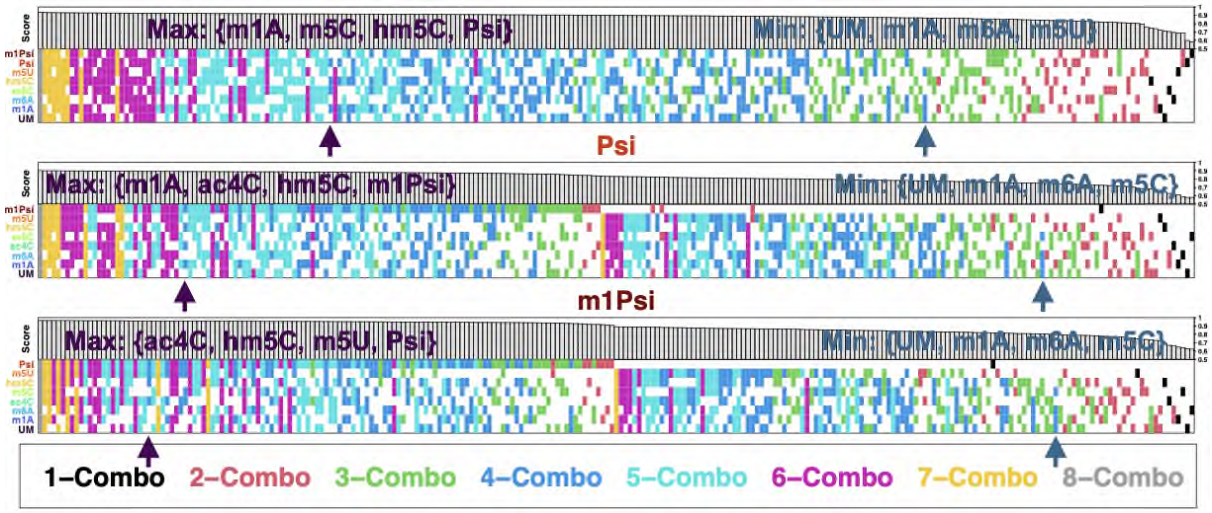


3. Analyses of the representation space are also descriptive, and the results are not surprising at all. The authors did not bother to discuss about why adding some modifications improved the basecalling accuracy but adding some other types did not.

In the updated manuscript, we systematically explored the question raised by the reviewer: “why adding some modifications improved the basecalling accuracy but adding some other types did not”. We explained such performance differences from the signal perspective, and reasoned that an ideal training modification combination will completely cover signals of the out-of-sample modification. Based on this hypothesis, we formally defined the “signal cover score”, as shown in the following figure:

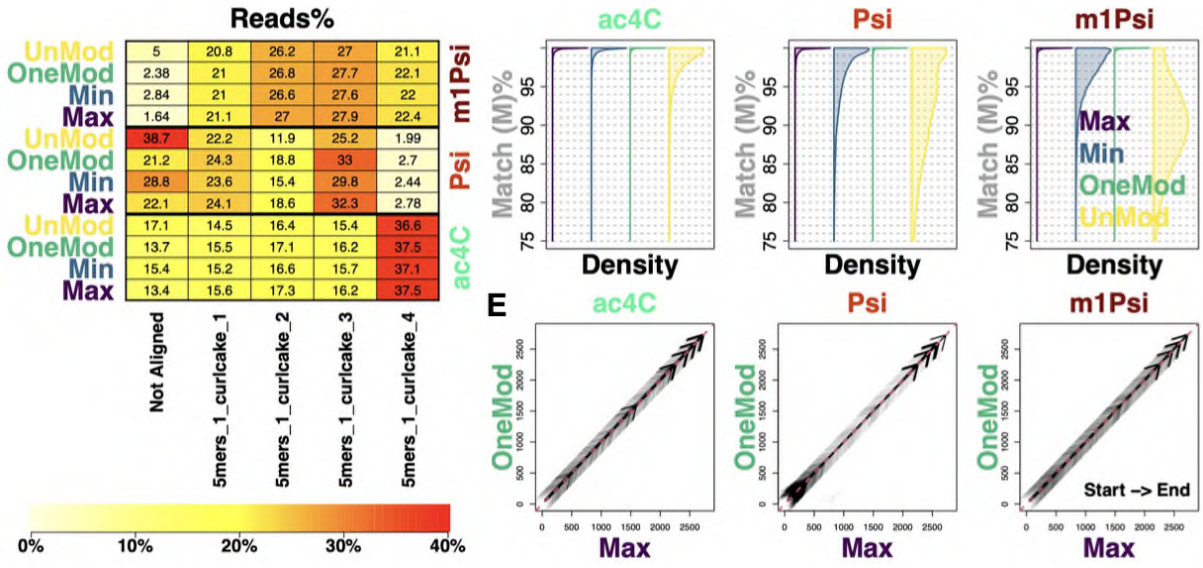


Based on such a score, for ac4C, Psi and m1Psi test groups, we evaluated all the possible training modification combinations, as shown in the following figure:



We would like to highlight that 1) combinations of more modifications produce higher scores, which echoes with our finding that diverse training data improves out-of-sample modification basecalling; 2) the inclusion of Psi and m1Psi training modification improve scores of m1Psi and Psi test data respectively, which echoes with our results that Psi and m1Psi reads could be acceptably handled by basecallers training using only m1Psi and Psi reads respectively.

We further systematically confirmed the signal cover score to be a metric to evaluate training combinations. Without losing generality, we analyzed all the 4-combos, and highlighted ones with the highest (Max) and lowest (Min) scores. As shown in the following figure, we noticed that “Max” out-performed “Min”, and accomplished comparable basecalling performance with “OneMod” positive controls (trained with only candidate modifications):



Altogether, we explained the reviewer’s question, by demonstrating the signal cover score as the metric for prioritizing training modification combinations. We reported these results in the updated Figure 3, as well as the RESULTS section “evaluating the out-of-sample basecalling generalizability for training modification combinations”:

"*We further asked, are all the training modifications required for accurate out-of-sample basecalling? To formally answer such a question, an evaluation metric for training modification combinations are required. Inspired by our findings that basecallers trained with only unmodified oligos basecalled certain modifications (m5C, hm5C, m5U) with acceptable accuracy (Figure 2A), as well as signals of such modifications were less deviated from their canonical counterpart (Figure S2, see METHODS), we hypothesized that the optimal training combinations can completely cover signals of the out-of-sample modification. We therefore defined the "signal cover score", with the following five steps: 1) correspond signal segments (events) with basecalled sequences; 2) assemble all the events mapped to the same sequence position, calculate their signal mean values, then use 10% and 90% quantiles (to avoid outliers) as the "effective signal range"; 3) quantify the training effective signal range, by taking the union of all the training modifications; 4) measure the cover fraction, as the test effective signal range that can be covered by the training counterpart; 5) calculate the final signal cover score, by taking the summation of all the cover fractions along the candidate sequence (Figure 3A).*

Based on such a signal cover score, for ac4C, Psi and m1Psi test groups, we evaluated all the possible training modification combinations. As shown in Figure 3B, we observed that combinations of more modifications in general produced higher signal cover scores, which echoes with our observation that diverse training data improves the basecalling of out-of-sample modifications. In particular, we noticed that the inclusion of Psi and m1Psi in training modifications significantly improved signal cover scores of m1Psi and Psi test data, respectively. These findings echo with our results that Psi and m1Psi reads can be acceptably handled by basecallers training using only m1Psi and Psi reads, respectively (Figure S4A and S5A). We further systematically confirmed that signal cover scores can be adopted to evaluate training modification combinations. Without losing generality, we prioritized the analyses of 4-combo combinations, and highlighted ones with the highest (Max) and lowest (Min) scores. We found that "Max" generated comparable mappability and basecalling accuracy with "OneMod" positive controls (basecallers trained with only candidate modifications), and out-performed "Min", particularly in Psi and m1Psi groups, as shown in Figure 3C and D. We finally confirmed that "Max" could generate consistent alignment with "OneMod" ground-truth, as shown in Figure 3E. Therefore, we concluded the signal cover score as the metric for prioritizing training modification combinations."

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The authors' revision has significantly improved the quality of the manuscript. However, some issues remain, particularly regarding the explanation of the method's generalization and the accommodation of new chemistries and models.

1. The authors retrained the basecalling model using the tRNA dataset to demonstrate improved tRNA basecalling accuracy. However, this improvement might result from learning specific k-mers within a small dataset. Moreover, this approach does not address the question: Can a model retrained from the curlcake dataset improve tRNA basecalling accuracy? The generalization capability of the trained model needs to be demonstrated through different approaches.

With all respect, we would like to explain that the tRNA analysis was used as an additional example (besides the curlcake analysis), to demonstrate that training data diversity enables the **modification-level generalizability** (precisely basecall of out-of-sample modifications). We therefore trained and compared, as described in the previous response letter, "positive" (with diverse modified tRNAs) and "negative" (with only non-modified tRNAs) models.

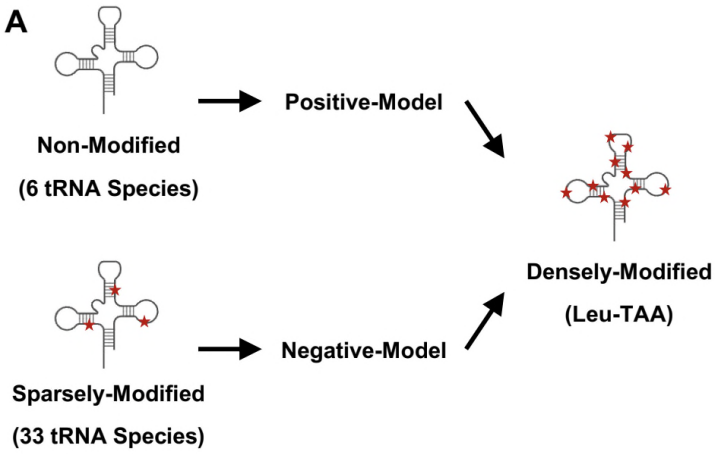
On the contrary, by asking "can a model retrained from the curlcake dataset improve tRNA basecalling accuracy", the reviewer indicated the **sequence context-level generalizability** (precisely basecall sequences with different kmer composition), which is out of the scope of this study. Indeed, the latest deep learning-based basecaller designs are context-aware, and are less generalizable to out-of-sample motifs. That explains when we analyzed tRNAs using the basecaller trained with the curlcake data, as suggested by the reviewer, the performance was significantly compromised.

Following the suggestion by the editor, we included the tRNA analysis, as we reported in the previous response letter (in response to your Comment #3 "can the author showcase if their retrained model performed better in any real datasets?"), in the revised manuscript.

Specifically, we included a new RESULTS section "Training data diversity enhances the basecalling of densely-modified tRNA reads":

"*Finally, we demonstrated the practical usefulness of our basecaller training paradigm by analyzing a yeast native tRNA nanopore sequencing dataset[16]. Specifically, we aimed to precisely basecall the most densely-modified Leu-TAA tRNAs, which contains 15 known modification sites. We further trained positive and negative-models, by combining the 33 sparsely-modified and 6 non-modified tRNA species, respectively (see METHODS). Our workflow was summarized in Figure 5A. We quantified the basecalling performance with mappability. We found that, compared to the negative-model which could only represent canonical sequences, the positive-model trained using diverse modifications achieved a ~15% increase in mappability (Figure 5B). Therefore, we highlighted our paradigm to be a general way of training modification-tolerant basecallers.*"

as well as the new Figure 5, in the updated manuscript:



B

	Not Aligned	Aligned
Positive-Model	716 (7.22%)	9,202 (92.78%)
Negative-Model	2,154 (21.72%)	7,764 (78.28%)

2. Much of the added discussion is somewhat shallow and computationally focused, lacking sufficient consideration of the underlying reasoning behind the findings. Please consider rewriting these sections to provide deeper insights. For example, in response to the observation that “deletion rates could be as high as 25%, while insertion and mismatch rates are 5–6%,” the authors stated “high-level signal deviations.” However, significant signal deviations would likely cause more mismatches along with indels, resulting in a proportional increase in both mismatches and indels. High indel rates may be related to changes in dwell time, but a higher rate of deletions rather than insertions could also be algorithm-related. There are many interesting details within these findings, but the authors have overlooked other possibilities without providing sufficient reasoning.

We thank the reviewer for the highly insightful comment, and fully acknowledge that the reviewer’s suggestion would make our manuscript more full-fledged. We therefore revised the DISCUSSION section accordingly:

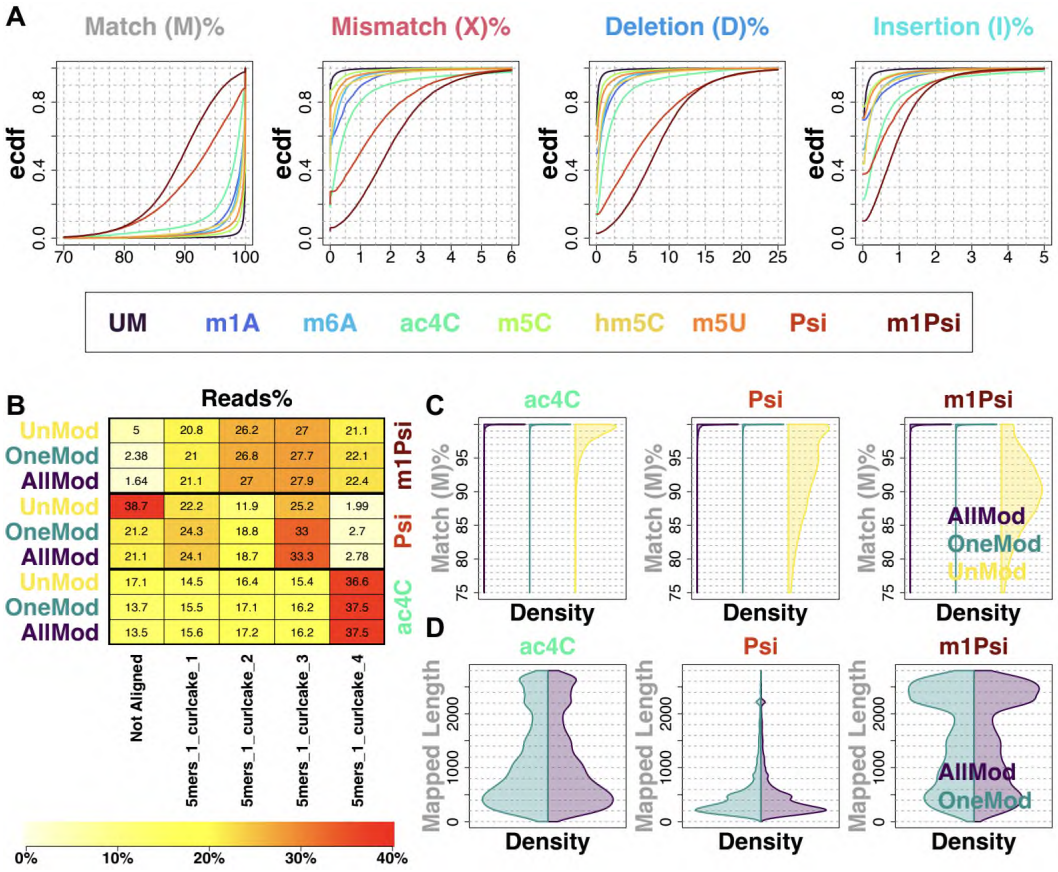
“*Promoting the basecalling accuracy, especially for native DNA/RNA sequencing signals, remains a central challenge in nanopore sequencing bioinformatics. This is because the latest basecallers are less tolerant to modifications, which commonly exist among native DNA and RNA molecules and will deviate nanopore sequencing signals. Specifically, we observed that shifted signals primarily cause mis-basecalls, while changes in dwell time mainly affect basecalled sequence lengths. For instance, for Psi and m1Psi, significantly deviated signals (quantified by mean and standard deviation) and dwell time (Figure S2) coincided with excessive mismatches and deletions/insertions, respectively (Figure 2A). As for ac4C, the presence of which particularly alters dwell time (Figure S2), we noticed insertions as the major type of basecalling artifacts (Figure 2A). To properly address this limitation, basecallers that are agnostic to nucleotide modifications are in urgent need.*”

Minor Point:

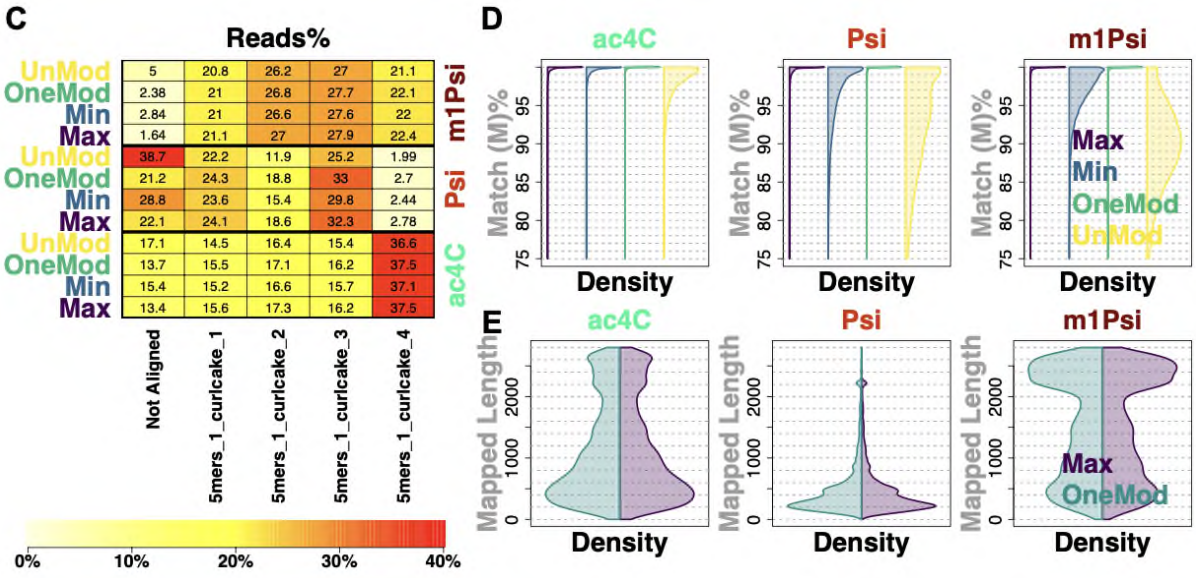
Consider replacing Figure 2D with two or more one-dimensional distributions (e.g., violin plots with boxplots) representing the mapped length ratio (0–100%). In the current figure, I can only discern the highlighted differences between X and Y.

We thank the reviewer for the comment, and updated Figure 2D (and the related Figure 3E and Figure S1) to show the mapped length distributions using violin plots.

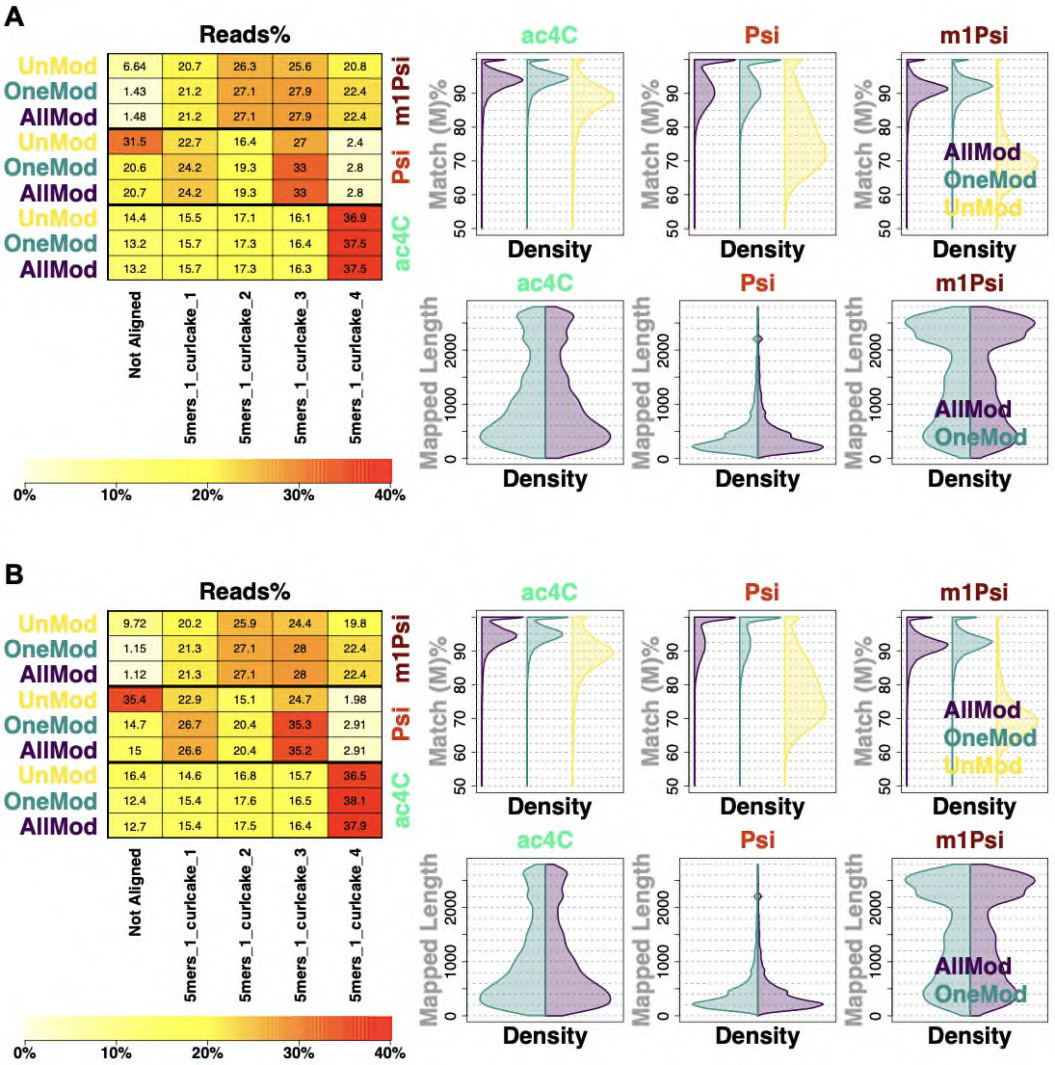
Updated Figure 2D:



Updated Figure 3E:



Updated Figure S1:



Reviewer #3 (Remarks to the Author):

I am afraid that my previous concerns have not been fully resolved. In the revised manuscript, the authors have added some more analyses with other pipelines. However, my concern lies in the methodology design and originality.

The major point of this study is that training of data with diverse modifications generalize the representation space, so that the basecalling for the nts with out-of-sample modifications would be more accurate. As I have mentioned in my previous comments, this idea itself is not original. Similar strategy has been used in other studies (e.g., the preprint on biorxiv, which presents a more novel method for detections of RNA modifications). Technical notes on ONT have also recommended such a strategy. I disagree with the authors on their claim that such training model has yet to be investigated. It is not “a novel paradigm for training modification-agonistic basecallers”.

The authors used different data to demonstrate that using diverse raw data for training can improve basecalling accuracy. This is empirical. Again, this work would be a nice technical note, but not a method paper with novel design or new concepts of methodology design. I simply don’t think it meets NC’s standard for innovation.

With all respect, we found that the comment from Reviewer #3 to be less constructive, and was probably made from misinterpretations of previous studies. Specifically:

Reviewer #3 cited a preprint (https://www.biorxiv.org/content/10.1101/2023.11.28.568965v1, which was recently published on Genome research) as an example of the “similar strategy” showing that our “idea itself is not original”.

However, this paper aims to **increase** modification-induced basecalling errors (as signatures for modification detection), which is exactly the **opposite** of our goal to **remove** errors for modification-tolerant basecalling. Specifically, in the abstract, the authors concluded that “*… alternative RNA basecalling models … increases the 'error' signal of m6A…*”, and “*… high-accuracy basecalling models … increased basecalling error signatures at pseudouridine (Ψ) and N1-methylpseudouridine (m1Ψ) modified sites*”. Therefore, we respectfully disagree with Reviewer #3 that this paper presents a “similar strategy”.

Reviewer #3 also mentioned that ONT technical notes “also recommended such a strategy”, to argue that our “idea itself is not original”.

Indeed, my lab just returned from the 2024 Nanopore Community Meeting. According to presentations by ONT employees, diverse training data was used to improve the basecalling of **known** modifications at known motifs (https://www.youtube.com/watch?v=rRFREd8qxgs, the example is for DNAs but they mentioned RNAs follow the same rationale). Our paper, on the other hand, focuses on the accurate basecalling of **previously-unseen**, **out-of-sample** modifications, which is conceptually innovative as opposed to the state-of-the-art.

In summary, we would like to re-emphasize that our paper demonstrated a novel paradigm for training modification-tolerant basecallers, which is the cutting-edge research direction (advocated by ONT during the Nanopore Community Meeting) in nanopore bioinformatics.