

DirectRM: Integrated Detection of Landscape and Crosstalk Between Multiple RNA Modifications Using Direct RNA Sequencing

Corresponding Author: Professor Kunqi Chen

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)

Yuxin Zhang and Yuecheng Wu et al. developed a new tool, DirectRM, to detect RNA modifications from direct RNA nanopore sequencing data. They aligned the ionic current data to the reference sequence (resquiggling) using Remora and then used a two-step classification process to first test if the position is likely modified and then followed with an attention model to determine the RNA modification identity. The first binary classification for presence of an RNA modification was done using a Wilcoxon test to measure the deviation of the mean ionic current intensity from the expectation and coupled this with basecaller errors at the position. The sites with likely modifications were classified using an attention model that uses the kmer sequence and features of the ionic current data, such as the dwell time, signal intensity mean, standard deviation, median, min, max, and quartiles to classify the site as one of six RNA modifications (m6A, m1A, m5C, m7G, pseudouridine, and ac4C). The authors calibrated and tested DirectRM using in vitro transcribed RNA, two human cell lines (HepAD38 and MT4), and human T-lymphotropic viral RNA cultured in MT4 cells. They also performed RNA modification-specific immunoprecipitation methods with NGS as an orthogonal measure of the RNA modifications and positive labels for DirectRM training. DirectRM has impressive RNA modification classification performance with AUC close to 0.9, however, they only match approximately 30% of the m6A-MeRIP-seq sites, however, this still outperforms the other tools they tested. The authors also detected neighboring RNA modifications and found that there was an antagonistic relationship between RNA modifications when within 100 nt of each other on the same molecules, except for m6A and m5C.

Overall, I was impressed with the rigor with which DirectRM was tested and impressed with the overall results. The paper is scientifically sound.

One critical aspect of the input data to the model is the resquiggling step. Did the authors try to use the established f5c eventalign or the relatively new uncalled4 as their resquiggler, which both have RNA002 and RNA004 functionality? I'm not aware of any published resquiggling performance tests for Remora other than it is what ONT is using for their RNA modification models. You should expand this section, ideally exploring resquiggling choices

One very minor note is that IVET was first defined on line 688 and not the first instance of the term in the text on line 125.

(Remarks on code availability)

Light review - great it was in github

Reviewer #2

(Remarks to the Author)

This work represents a groundbreaking advancement in epitranscriptomics, providing integrated analysis of multiple RNA modifications. The author's development of DirectRM, a computational framework leveraging nanopore direct RNA sequencing and deep learning, is both technically innovative and biologically insightful.

A major strength of this work lies in its use of natural RNA samples for both feature extraction and label derivation, ensuring that the model capture the biological relevance, which is absent in IVT samples. The model was trained with nanopore data and NGS data generated from the same cell lines, which mitigate cross-platform biases and ensure the NGS-label reflect true modification landscapes. This alignment is critical, as modifications exhibit cell-type-specific patterns.

Its two-stage detection pipeline elegantly overcomes the limitations of existing single modification detection tools. By combining de novo kmer screening with attention-based multi-instance multi-label model, the authors achieved simultaneous detection of six RNA modifications. The reduction in false positive rate compared to existing tools is particularly compelling. Also, the decision to optimize sub-model structure with m6A, which has well-defined motifs, demonstrates the methodological rigor.

The authors provide rigorous validation across multiple axes, such as validation with NGS techniques, tested the biological reproducibility under different biological contexts, benchmarking with existing modification detection tools, as well as experimental validation of DirectRM results on viral RNAs.

The read-level and transcript-level exploration of crosstalk between RNA modifications is interesting, providing meaningful insight into the complex epitranscriptome. This duality is further enriched by the linkage of RNA modifications with RNA binding proteins.

Its ability to profile "dark" regions of the epitranscriptome (e.g., rRNAs, low-abundance isoforms) addresses a major limitation of NGS-based approaches.

This manuscript provides valuable resources by generating and publicly releasing direct RNA sequencing data and paired NGS data across multiple cell lines, which is unprecedented in scale and diversity for future RNA modification research.

Below are some concerns and points that need to be addressed before publication:

1. When optimizing the de novo detection model, the author tried three indicators. The authors chose the Wilcoxon test to measure the difference between current intensities and expected levels. The authors should clearly explain the reason of using Wilcoxon test rather than other tests. Also, it seems that the sample size for the test is not sufficient ($n = 5$ or 9), please clarify the reliability of test results. Regarding the base call error features, four base call features were extracted, but only insertion, deletion or mismatches were used (4.2.2), why not consider base with decreased base call qualities?
2. The authors stated that 30x coverage cutoff was used. Figure 1C shows the relationship between sensitivity and coverage cutoff. I see 10x coverage can provide satisfying sensitivity, using 30x coverage may ignore transcript with low expression. Also, the authors could clarify how coverage affect the detection specificity.
3. The sub-model structure was optimized for m6A due to its well-characterized motifs and distributions. Could this introduce m6A-related biases when generalizing to other modifications (e.g., ac4C with less defined motifs).
4. During the feature extraction process, the authors used a 9mer sliding window and default step size of 5. This may lead to partial overlaps between adjacent instances, individual positions may be presented in multiple instances. The authors should clarify how different measurements for the same position integrated, how conflicting predictions from overlapping instances are resolved?
5. The choice of window size (9-mer) and step size (5 nt) may bias detection toward modifications in specific sequence contexts. Please discuss this point.
6. The observation that adjacent modifications "reduce" each other's occurrence on the same molecule is intriguing. Could this reflect steric hindrance during sequencing (e.g., pore occupancy artifacts) rather than biological competition?
7. The 3,727 positive associations with RNA-binding proteins (RBPs) are noted, but are these RBPs known to interact with the specific modifications detected (e.g., m6A readers like YTHDF2)? A deeper analysis linking RBP motifs to modification pairs would strengthen claims.
8. The comparison between HepAD38 and MT4 cells is underdeveloped. Are the observed differences in modification rates attributable to cell lineage (e.g., hepatocyte vs. T-cell biology) or technical variability?
9. When measuring the distance between modification sites, only linear distance was considered. Could the analysis integrate RNA secondary structure features or for circular RNAs, the back splicing features, when measuring the distance between sites.
10. This manuscript analyzed the association between modifications on HepAD38 cell lines. Since MT4 samples are also available, the authors could perform similar analysis on MT4 cell lines as well, and see whether there is cell-line-specific patterns.
11. The reliance on NGS-based labels for model training limits applicability to modifications without established profiling methods. Could synthetic RNAs with defined modifications supplement training data?
12. In terms of the software usage, provides details on the scalability of software. Can DirectRM handle large-scale datasets typical of nanopore sequencing?

Below are minor edits and clarifications for the manuscript:

1. The manuscript introduces abbreviations without defining them at first mention, which may hinder readability for non-specialist audiences. Examples include IVET in line 125.
2. The manuscript occasionally uses different abbreviations for the same metric, which may confuse readers. For example: AUC (Area Under the Curve) and AUROC (Area Under the Receiver Operating Characteristic Curve) are used interchangeably (e.g., line 129 vs. line 124). While both refer to the same concept, consistency is critical for clarity.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

The paper by Zhang et al studies the crosstalk between RNA modifications with a modified multiple instance learning algorithms, with 3 steps in the module, 1) de novo detection of modified kmer sites, 2) a binary classification layer for 6 sub-models with each model for one modification type to extract attention-scores for different features, 3) a optimization step that learn integratively multi-labels. Through learning and predicting for multiple modification sites simultaneously, the authors further found that within 100bp, modification level is lower the closer modification site neighbors, and of the modification site neighbor pairs, by comparing marginal probability against conditional probability, the authors showed the positive and negative impact between pairs of modification types, and identified similar results as reported previously. The study is generally interesting and worth reading, however, there are a number of critical concerns regarding the method evaluation and analysis, as well as results presentation. Please see my comments below for more details:

Major comments

While the concept of the method is interesting, the overall presentation of DirectRM is confusing and distractive, without clear focus. For example, section 2.1 and 2.2, are less interesting and similar to existing approaches, can be reduced, while 2.4 where comparing independent model and integrative model optimization should be highlighted more. Related to that, the modification position determination results seems unnecessary, suggest to remove this from the manuscript.

The evaluation and benchmarking of the method is not convincing. For each module of the method, currently, it seems rather arbitrary, and comparison is within the method itself, hence hard to justify why this is good and why this is better than existing approaches. Suggest to include other methods, when benchmarking each of the steps, and also when applying to different cell line dataset. Another point is that not many datasets are evaluated here, would suggest adding more datasets in benchmarking, can be public datasets, or new datasets generated, or different species datasets.

When motivating for the simultaneous modelling of multiple RNA modifications, the description is very unclear, suggest to rephrase this part, to highlight why need to joint model, and why independent model is not working. Currently, the description is limited to remain false positive issues, but false positive issue is a general problem that is with every method, so this is not a very convincing point. IVT RNA-based training would distort the model because of unnatural residues, also needs more support as to how often IVT RNA would have unnatural residues, if not often, then probably this also reflects the true scenario where consecutive modifications would happen.

Can the authors elaborate how the multi-label learning is done, what are the task and how is this different from single-label learning? What is the main task in the multi-label learning step? The current description in section 4.2.3 is very limited and hard to comprehend.

Minor comments:

Lines 175-177: The results of comparing the contribution from sequence features against signal and baseball error features are not well analyzed. The conclusion of the sequence feature is more significant seems only applicable to m6A based on Figure 2A and Figure S1. Firstly, there is no test associated with this conclusion. Secondly, based on the figure itself, this is mainly based on sum of each type of features, if all features compared together with a ranking, then it probably can be seen that one or two of the sequence position stands out as very important, with some of the features from signal and baseball error stands out. Would suggest this be reorganised and the results refined.

Line 196: average AUROC>0.9, but Figure 2B all below 0.9, the statement needs to be refined or changed, as it is not reflecting the truth based on the figure.

Lines 201-203: The results presented in Figure 2C are confusing. In the first place, why evaluate % of DRACH motif against all m6A sites? Should the absolute number be evaluated for DRACH and total m6A sites separately? Because the % of DRACH is affected by the total number of identified m6A sites. Even % of DRACH is high, but the number of DRACH is low, then it is not considered good, I think.

Line 203-205: results on Figure 2D also need better clarification, why is there a second mode in the end around the end of 3' UTR in all models, but not NGS background profile, this require more discussion and investigation, whether this is related to systematic error in model labels or model specification? Or is it there in the training dataset?

For Figure 3D, the integrated model detects much less significant sites, are they false positives? Could there be true positives missed by the integrated model? This is not evaluated and is worth exploring before a conclusion can be made. Similarly, for lines 247 to 253, this is worth more analysis. Currently, this is all speculation. It would be great if this could be evaluated with some kind of ground truth label dataset. It probably can be tested with some other synthetic datasets, with a mixture of different degrees of multiple modifications present within one kmer.

Line 80: typo "there still limitations"?

Line 81: psi, better to use the complete word instead?

Line 81-82: "only report non-canonical bases without specifying their modification types...", need citation specifically to make the statement more clear, currently unclear what the author is trying to say

Line 87-88: this sentence describing methods being different, and hence hard to collate and combine results, the statement need to be rephrased, maybe the author mean the differences in method makes results highly specific to each method, and hence hard to compare? But what is this statement leading to? Need more clarification

Line 152: "coverage." should be "coverage cutoff."?

Line 153: with increased coverage cutoff, more modified sites are identified. Because, absolute number of sites with coverage cutoff is reducing, it is the distribution of modified vs unmodified sites become larger. Suggest to rephrase.

Line 207: typo, "is correlate to"

Figure 3 Legend: missing (E) legend

Line 331: tissues-specific?

Line 336, Figure 4F, should be 5F?

Line 338: What does "the dynamic feature RNA modification" mean?

Line 373: Tax isoform?

Line 397: The sentence is incomplete

Line 404 Figure 7B: p-value is not provided for the comparison

(Remarks on code availability)

I have not personally run the code, because of lacking of test dataset. Suggest the authors to include a small test dataset for users to try out.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors of DirectRM have improved the manuscript and addressed our previous concerns. The benchmarking between Remora and Uncalled4 revealed that the resquiggling algorithm did not have a substantial impact on the RNA modification classification precision or recall. This flexibility allows others to integrate DirectRM to existing direct RNA nanopore sequencing data processing pipelines.

Github review:

I do not necessarily think these are required but would be helpful to increase the adoption of DirectRM by the community.

The github page specifies that Dorado v0.6 is required for DirectRM. Is there anything specific to Dorado v0.6 that is required for DirectRM, or is that the only Dorado version that you tested? ONT released Dorado v1.0 this summer, and they usually update their basecalling software once or twice a year. Dorado v1.0 has increased accuracy and read lengths compared to earlier Dorado versions. DirectRM will stay relevant in the coming years if it is compatible with current Dorado versions.

It would be helpful to include a requirements.txt file, a conda environment yml file, a docker or singularity image, or some method to ensure all the required python libraries are properly installed when using DirectRM.

(Remarks on code availability)

See above

Reviewer #2

(Remarks to the Author)

I have no further comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

The code was well organized in github and I successfully installed it and got it to run on the test data set provided.

Reviewer #4

(Remarks to the Author)

The authors have satisfactorily addressed all my concerns, and I greatly value their hard work.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Yuxin Zhang and Yuecheng Wu et al. developed a new tool, DirectRM, to detect RNA modifications from direct RNA nanopore sequencing data. They aligned the ionic current data to the reference sequence (resquiggling) using Remora and then used a two-step classification process to first test if the position is likely modified and then followed with an attention model to determine the RNA modification identity. The first binary classification for presence of an RNA modification was done using a Wilcoxon test to measure the deviation of the mean ionic current intensity from the expectation and coupled this with basecaller errors at the position. The sites with likely modifications were classified using an attention model that uses the kmer sequence and features of the ionic current data, such as the dwell time, signal intensity mean, standard deviation, median, min, max, and quartiles to classify the site as one of six RNA modifications (m6A, m1A, m5C, m7G, pseudouridine, and ac4C). The authors calibrated and tested DirectRM using in vitro transcribed RNA, two human cell lines (HepAD38 and MT4), and human T-lymphotropic viral RNA cultured in MT4 cells. They also performed RNA modification-specific immunoprecipitation methods with NGS as an orthogonal measure of the RNA modifications and positive labels for DirectRM training. DirectRM has impressive RNA modification classification performance with AUC close to 0.9, however, they only match approximately 30% of the m6A-MeRIP-seq sites, however, this still outperforms the other tools they tested. The authors also detected neighboring RNA modifications and found that there was an antagonistic relationship between RNA modifications when within 100 nt of each other on the same molecules, except for m6A and m5C.

Overall, I was impressed with the rigor with which DirectRM was tested and impressed with the overall results. The paper is scientifically sound.

Re: We thank the reviewer for acknowledging the rigor of DirectRM. We have addressed all subsequent comments point-by-point below to further strengthen the manuscript.

One critical aspect of the input data to the model is the resquiggling step. Did the authors try to use the established f5c eventalign or the relatively new uncalled4 as their resquiggler, which both have RNA002 and RNA004 functionality? I'm not aware of any published resquiggling performance tests for Remora other than it is what ONT is using for their RNA modification models. You should expand this section, ideally exploring resquiggling choices

Re: We thank the reviewer for this insightful comment. We fully agree that benchmarking against resquiggler is critical for contextualizing our work. In the revised manuscript, we added a new section named "**Benchmarking on resquiggle algorithms**" to the **Supplementary file**.

In the revised version, we carefully compared the Remora with Uncalled4 as the resquiggler. In terms of f5c, since it adopted distinct input and output data formats, and was reported to have relatively poorer performance than Uncalled4, we excluded it from the comparison¹.

The training HepAD38 samples, sequenced with both RNA002 and RNA004 chemistry, were used for comparison. We trained the multi-label models with data processed with Uncalled4 and Remora, and evaluated their performance with multi-label accuracy, multi-label PRAUC, and AUROC for each class. **Table R1** summarizes the performance of models based on different resquiggle algorithms. The results suggested that Remora showed slightly better performance than Uncalled4 on both RNA002 and RNA004 chemistry. Thus, we decided to remain using the Remora as the resquiggle algorithm.

While our primary pipeline integrates Remora as the resquiggler, we also maintained the compatibility with Uncalled4 to ensure the flexibility for users with existing Uncalled4 workflows. Guidance for integrating Uncalled4 to the DirectRM framework can be found on the revised GitHub guidance page.

Table R1. Performance of models based on different resquiggler.

		multi-ACC	multi-PRAUC	ac4C	m1a	m5C	m6A	m7G	Ψ
RNA002	Remora	0.798	0.931	0.962	0.972	0.977	0.973	0.976	0.979
	Uncalled4	0.784	0.926	0.962	0.970	0.976	0.972	0.974	0.977
RNA004	Remora	0.795	0.929	0.963	0.972	0.977	0.973	0.975	0.978
	Uncalled4	0.793	0.928	0.961	0.971	0.978	0.973	0.975	0.978

One very minor note is that IVET was first defined on line 688 and not the first instance of the term in the text on line 125.

Re: we thank the reviewer for catching this oversight. The term “IVET” has now been formally defined at its first occurrence in line 138 in the revised manuscript. In addition, we have carefully proofread the manuscript to ensure that all acronyms now strictly follow the academic formatting instructions.

Reviewer #1 (Remarks on code availability):

Light review - great it was in github

Re: we thank the reviewer for this comment.

Reviewer #2 (Remarks to the Author):

This work represents a groundbreaking advancement in epitranscriptomics, providing integrated analysis of multiple RNA modifications. The author’s development of DirectRM, a computational

framework leveraging nanopore direct RNA sequencing and deep learning, is both technically innovative and biologically insightful.

A major strength of this work lies in its use of natural RNA samples for both feature extraction and label derivation, ensuring that the model capture the biological relevance, which is absent in IVT samples. The model was trained with nanopore data and NGS data generated from the same cell lines, which mitigate cross-platform biases and ensure the NGS-label reflect true modification landscapes. This alignment is critical, as modifications exhibit cell-type-specific patterns.

Its two-stage detection pipeline elegantly overcomes the limitations of existing single modification detection tools. By combining de novo kmer screening with attention-based multi-instance multi-label model, the authors achieved simultaneous detection of six RNA modifications. The reduction in false positive rate compared to existing tools is particularly compelling. Also, the decision to optimize sub-model structure with m6A, which has well-defined motifs, demonstrates the methodological rigor.

The authors provide rigorous validation across multiple axes, such as validation with NGS techniques, tested the biological reproducibility under different biological contexts, benchmarking with existing modification detection tools, as well as experimental validation of DirectRM results on viral RNAs.

The read-level and transcript-level exploration of crosstalk between RNA modifications is interesting, providing meaningful insight into the complex epitranscriptome. This duality is further enriched by the linkage of RNA modifications with RNA binding proteins.

Its ability to profile "dark" regions of the epitranscriptome (e.g., rRNAs, low-abundance isoforms) addresses a major limitation of NGS-based approaches.

This manuscript provides valuable resources by generating and publicly releasing direct RNA sequencing data and paired NGS data across multiple cell lines, which is unprecedented in scale and diversity for future RNA modification research.

Re: We are grateful for the recognition of our work. We have addressed all subsequent comments point-by-point below to further strengthen the manuscript.

Below are some concerns and points that need to be addressed before publication:

1. When optimizing the de novo detection model, the author tried three indicators. The authors chose the Wilcoxon test to measure the difference between current intensities and expected levels. The authors should clearly explain the reason of using Wilcoxon test rather than other tests. Also, it seems that the sample size for the test is not sufficient ($n = 5$ or 9), please clarify the reliability of test results. Regarding the base call error features, four base call features were extracted, but only insertion, deletion or mismatches were used (4.2.2), why not consider base with decreased base call qualities?

Re: We thank the reviewer for these insightful comments.

Regarding the choice of the Wilcoxon test, it was based on the following considerations:

- 1) The current intensities do not follow a normal distribution. As a non-parametric statistical method, the Wilcoxon test is particularly suited for such a case.

- 2) Our experimental setup compared the observed value with its expected level at identical kmer positions, constituting matched pairs. The Wilcoxon test explicitly models this dependency.
- 3) Wilcoxon test, as a non-parametric test, is suitable for samples with a small sample size.
- 4) The Wilcoxon test is used by the ONT Tombo software to measure the discrepancies between observed and expected signals.

Therefore, we decided to use the Wilcoxon test, as it is suitable for the electric signal features.

Regarding the reliability of test results on small sample sizes, firstly, the chosen Wilcoxon test is powerful for samples with a small sample size. The excellent performance of trained models (AUROC > 0.9) also indicated the reliability of the test results.

In terms of base call error features, we fully agree that including the “decreased base call qualities” may benefit the identification of modified kmers. In the revised model, in addition to the presence of insertion, deletion, and mismatches, the presence of bases with decreased qualities is also used as an indicator of modifications. In order to find an optimal cutoff for decreased base call qualities, we used the IVET dataset. **Figure R1** illustrates the distribution of base qualities for the modified and unmodified datasets. The result aligns with the expectation that the modified dataset has significantly lower base qualities. Then, we use the ROC curve optimization method to find the optimal cutoff. The Youden’s J statistic (TPR-FPR) across all possible thresholds was calculated, and the values that maximize the metrics were selected as the cutoff: base call quality < -1. The identification of the optimal threshold for decreased base qualities can be found in the **Supplementary file**, which is named “**Identification of optimal threshold for decreased base qualities**”.

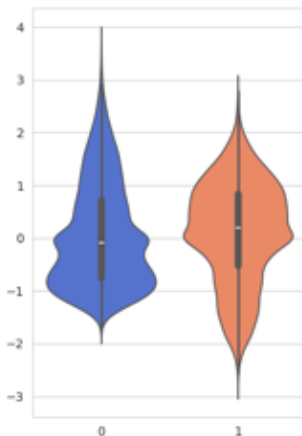


Figure R1. Distribution of base call qualities on modified and unmodified datasets.

2. The authors stated that 30x coverage cutoff was used. Figure 1C shows the relationship between sensitivity and coverage cutoff. I see 10x coverage can provide satisfying sensitivity, using 30x coverage may ignore transcript with low expression. Also, the authors could clarify how coverage affect the detection specificity.

Re: We thank the reviewer for these insightful comments. We systematically evaluated the influence of coverage on both the Recall (similar to sensitivity) and Precision (similar to specificity) with the IVET datasets. This part is included in the **third paragraph** of the **Strict labeling enables high-accuracy neural de novo modification detection** section under the **Results and Discussion** part and was highlighted.

According to **Figure R2**, as the coverage increases, the recall score increases significantly, indicating that more modification sites can be identified, which is consistent with previous findings. In the meantime, the Precision score decreased slightly, suggesting potential false positives.

To find a tradeoff between precision and recall, we calculated the F1-scores under different coverage. As the coverage increased from 1x to 10x, the F1-score improved sharply from 0.45 to 0.69, and then stabilized or slightly decreased beyond 10x coverage. This indicates 10x coverage represents a performance inflection point, where additional read coverage does not necessarily improve the performance and may ignore transcripts with low expression. Therefore, we decided to use 10x coverage in the revised version.

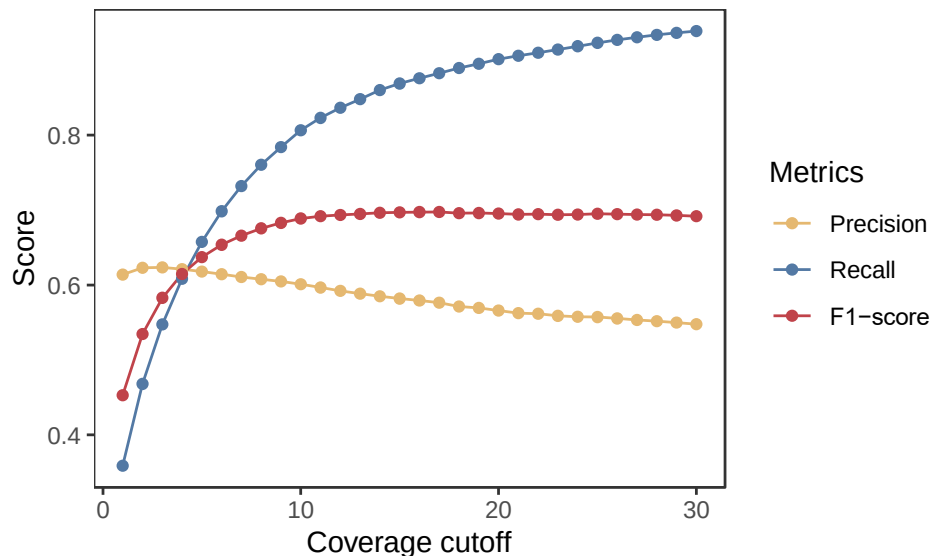


Figure R2. Precision, recall, and F1 scores under different coverage cutoffs.

3. The sub-model structure was optimized for m6A due to its well-characterized motifs and distributions. Could this introduce m6A-related biases when generalizing to other modifications (e.g., ac4C with less defined motifs).

Re: Thank you for raising this important concern. We acknowledge the need to clarify how our model architecture selection process avoids introducing m6A-specific bias to other modifications. Below, we address this systematically:

1. Our study primarily defined four distinct sub-model architectures, all achieving excellent performance on each modification (AUROC > 0.85). Since m6A has well-annotated biological features (DRACH motifs and transcriptomic distributions), it was used to evaluate

the model's ability to learn biological patterns, which is a critical requirement for biological interpretability. The selection process focused on the general learning ability of models rather than encoding m6A-specific inductive biases.

2. Critically, each modification's submodel operates with fully independent parameters: the attention mechanisms, feature extractor, and dense layers are not shared across modifications.
3. While the architecture was selected for its ability to learn m6A's DRACH motifs and transcriptomic distributions, this does not force other sub-models to adopt similar strategies. For modifications without conserved motifs, the same architecture learns distinct detection principles.

4. During the feature extraction process, the authors used a 9mer sliding window and default step size of 5. This may lead to partial overlaps between adjacent instances, individual positions may be presented in multiple instances. The authors should clarify how different measurements for the same position integrated, how conflicting predictions from overlapping instances are resolved?

Re: Thank you for highlighting this important technical consideration.

We used the 9mer sliding window with a step size of 5 during the feature extraction process, which creates 4-positional overlaps between consecutive instances. The reason is to ensure critical motifs (e.g, DRACH for m6A) are fully captured even at transcript ends, providing context redundancy to mitigate edge effects. By doing this, individual positions may have multiple estimations, and results (i.e., probability) will be averaged to provide the final modification probability. Clarification for this point was included in the revised manuscript in the last paragraph in the “**Modification type and exact location inference**” section of the **Materials and Methods** section.

5. The choice of window size (9-mer) and step size (5 nt) may bias detection toward modifications in specific sequence contexts. Please discuss this point.

Re: We appreciate the reviewer's insightful comment regarding the bias related to the sliding window size and step size. We fully acknowledge that the window size and step size selection can influence the sensitivity or specificity of modification detection. Below, we provide a detailed discussion of this point:

- **The choice of a 9-mer window size** is fundamentally tied to the biophysics of nanopore direct RNA sequencing. In the RNA004 chemistry, the nanopore can simultaneously harbor 9 nucleotides. The ionic current signal at any given time point directly reflects the composite physicochemical properties of the 9-mer occupying the pore. Our 9-mer windowing strategy is therefore biophysically grounded to ensure 1:1 alignment between extracted features and raw signal segments.
- **A 5-bp step size** ensures 50% overlap between consecutive windows, which:
 - Increases feature density in regions with ambiguous signals

- Mitigates edge effects by allowing modifications near window boundaries to be captured by multiple overlapping windows. For example, the critical motifs (DRACH) for m6A can be fully captured even at the end of the window.
- A step size smaller than 5bp (e.g., 1bp) would generate a proportional increase in the number of instances, leading to exponential growth in memory and storage usage. Thus, it would render large-scale analyses.
- In terms of the bias detection toward modifications in specific sequence context, we acknowledge that a fixed 9-mer window may underrepresent modifications on longer sequence motifs, and 2) a 5-bp step size might systematically oversample certain motifs or sequence context.
 - Regarding the first point, we agree that integrating longer sequence features could improve the detection for certain modifications. However, our method does not solely depend on sequence motifs for modification detection. Instead, it uses the raw current signals reflecting physicochemical distortions caused by modifications within the 9-mer pore window.
 - Regarding the second point, we calculated the frequency of 9-mers extracted with different step sizes:
 - step size = 5, 50 overlaps
 - step size = 9, no overlaps

It was found that their distribution is aligned, suggesting that the 5-bp step size is not likely to oversample certain motifs or sequence context (**Figure R3**).

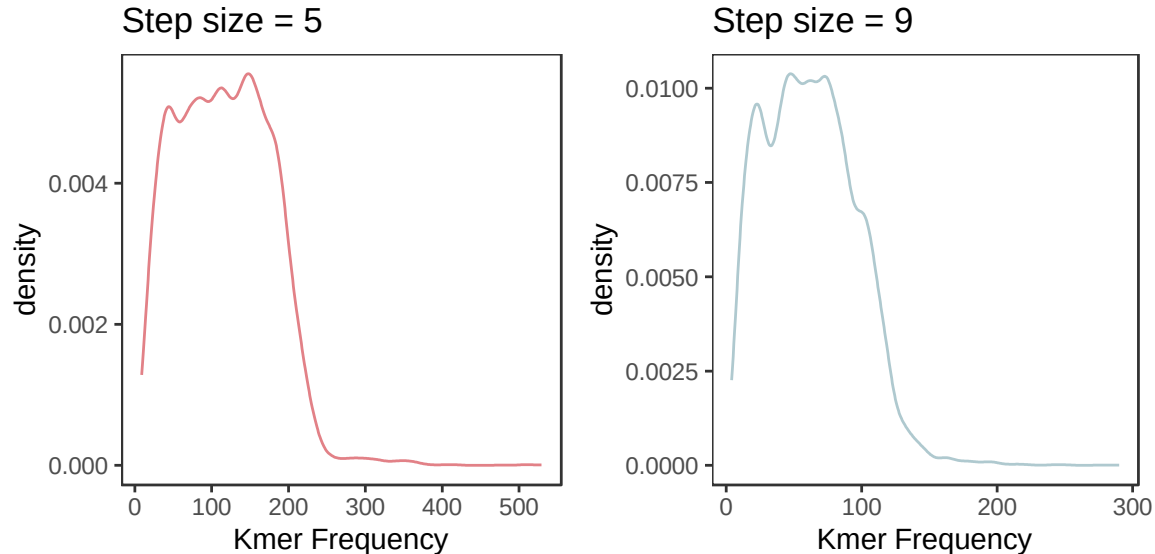


Figure R3. Kmer frequency with different step sizes.

In conclusion, the choice of a 9-mer window and a 5-bp size is biophysically grounded and empirically validated. We hope the above explanations resolve your concerns.

6. The observation that adjacent modifications "reduce" each other's occurrence on the same

molecule is intriguing. Could this reflect steric hindrance during sequencing (e.g., pore occupancy artifacts) rather than biological competition?

Re: We thank the reviewer for raising this critical consideration. We have carefully considered both technical artifacts and biological explanations. Below, we provide a detailed discussion of this point:

1. We agree that nanopore sequencing can theoretically introduce local steric effects, as the pore can accommodate 9 nucleotides at the same time. The presence of multiple modifications within the nanopore may cause unexpected errors.
2. However, the observation persists at distances far beyond 9bp. According to **Figure R4**, we evaluated the relationship between site modification rate and the distance to the nearest modification site. For sites that do not have modified neighbors within 9bp, but have modified neighbors beyond 9bp (10- 90bp), their modification rate is also lower than others. This suggests the phenomenon is unlikely to stem purely from pore occupancy artifacts, as such technical biases would disappear beyond the 9-mer window.
3. Previous studies, including the work of Acera Mateos, P. et al. ², and the work of Liu, C. et al ³, also reported similar observations, that the presence of one modification will influence the presence of nearby modifications.

In conclusion, we believe that this observation is biologically grounded. We hope the above explanations resolve your concerns.

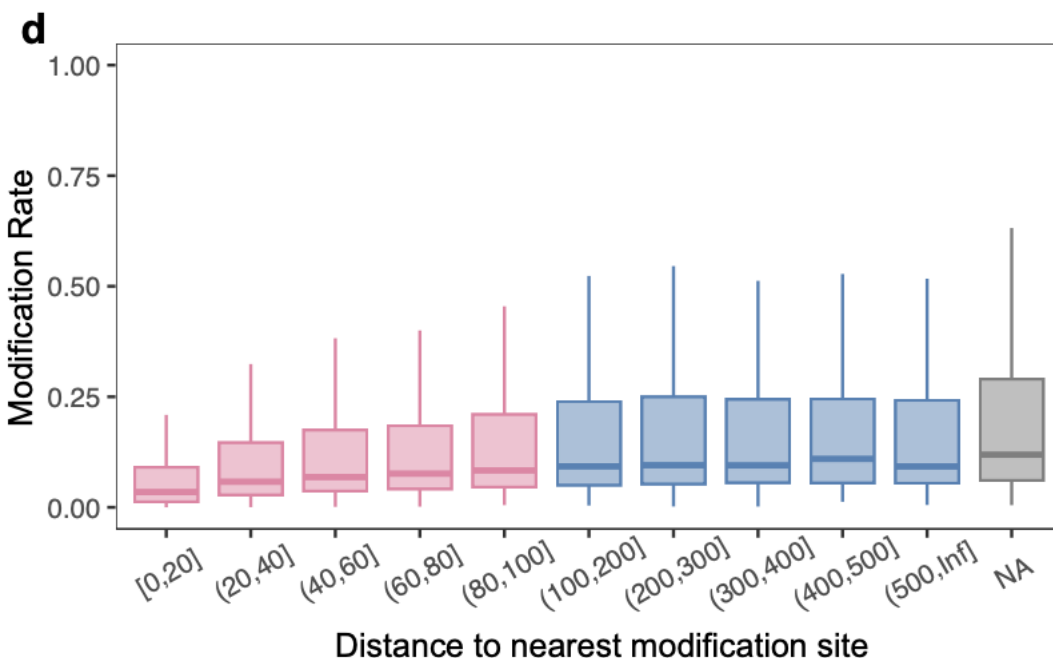


Figure R4. Crosstalk between modifications (d) Modification rate for sites grouped by distance to their nearest modification sites. NA indicates that the site is the only detected modification on that transcript.

7. The 3,727 positive associations with RNA-binding proteins (RBPs) are noted, but are these

RBP known to interact with the specific modifications detected (e.g., m6A readers like YTHDF2)? A deeper analysis linking RBP motifs to modification pairs would strengthen claims.

Re: We thank the reviewer for this insightful comment regarding the RBP and modification associations.

Current linkages are primarily based on transcriptomic coordinate overlaps between modification sites and RBP binding regions. While this approach provides a systematic initial survey of potential interactions, we fully acknowledge that this does not necessarily indicate functional engagement: modifications may reside within protein-binding domains without direct biological interplay.

We agree that integrating RBP motif specificity or binding preference (e.g., YTHDF2 for m6A modifications) would significantly strengthen the claims. However, this faces substantial data constraints: a very limited number of RBPs (e.g., YTHDF2 and IGF2BP3) have experimentally-characterized associations to specific modification or modification pairs.

Where possible, the three examples used for demonstration in the **“Molecule-level investigation revealed potential association between modifications”** in the **Results and Discussion** part are experimentally validated. We also discussed this limitation in the **Conclusion** part (highlighted) in the revised manuscript.

8. The comparison between HepAD38 and MT4 cells is underdeveloped. Are the observed differences in modification rates attributable to cell lineage (e.g., hepatocyte vs. T-cell biology) or technical variability?

Re: We appreciate the reviewer’s critical assessment of the HepAD38 and MT-4 comparison. While the technical variability was rigorously controlled:

- Identical library preparation, sequencing depth, and analysis pipelines were used
- Direct RNA sequencing adopted a simplified and straightforward sequencing pipeline, which significantly reduced batch effects and artifacts, providing higher reproducibility

Therefore, we expected that the observed differences in RNA modification profiles likely reflect cell-specific biology rather than artifacts.

Furthermore, in the revised manuscript, we performed comparisons between both cell lines (HepAD38 vs MT-4) and replicates (MT-4 rep 1 vs MT-4 rep 2). According to **Figure R5f**, while dynamic modification patterns can be found between replicates, a more significant difference can be identified between different cell lines. In addition, the pairwise comparisons demonstrated significant divergence in the modification rate between the cell lines. In contrast, site modification rates remained relatively static across the two MT-4 replicates (**Figure R5g**). This suggests that the observed difference is attributable to cell lineage rather than technical variability.

In addition, RNA modification is known to be a dynamic and reversible process; thus, the observed difference is consistent with RNA biology. We hope the above explanations resolve your concerns.

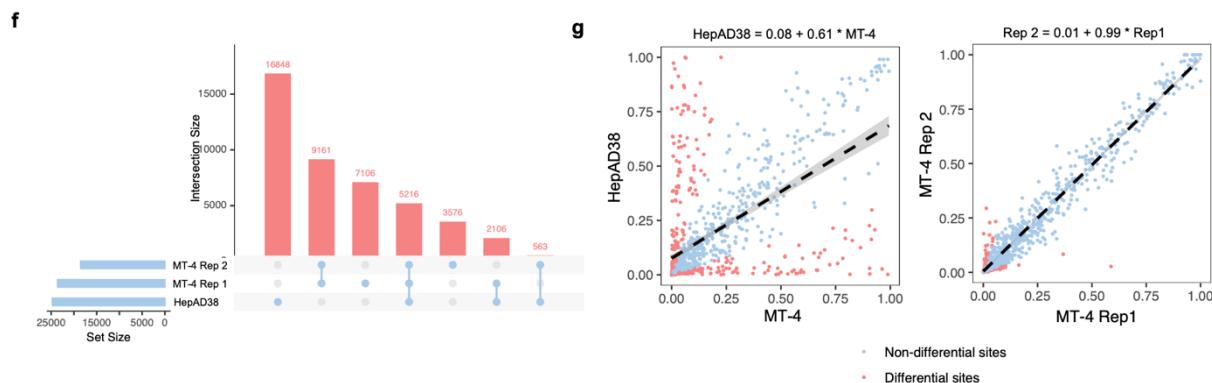


Figure R5. Identification of modifications on the human MT-4 cell line. (f) Upset plot showing the intersection of HepAD38 samples and the two replicates of MT-4 samples. (g) Dot plot showing the site modification rate under HepAD38 cell lines, and two replicates of MT-4 samples. Differentially modified sites under different cell conditions or replicates were defined as: fold change smaller than 1/3 or greater than 3.

9. When measuring the distance between modification sites, only linear distance was considered. Could the analysis integrate RNA secondary structure features or for circular RNAs, the back splicing features, when measuring the distance between sites.

Re: We appreciate the reviewer for this exceptionally insightful suggestion. We acknowledge that integrating RNA structural contexts (e.g., secondary structure or circularization) fundamentally addresses a critical dimension in epitranscriptome analysis, where the RNA structure and modification are interconnected to each other. However, we may fail to implement it in the current study due to technical constraints:

1. RNA secondary structure determination:

- Experimental determination of RNA secondary structure is laborious and low-throughput
- Based on our previous studies, existing computational prediction tools (e.g., RNAstructure⁴ and ViennaRNA⁵) exhibit substantial inconsistency in secondary structure predictions, particularly for long transcripts, suggesting the poor reliability of computational predictions. This methodological limitation would introduce uncontrolled artifacts into distance calculation and downstream analysis.

2. Circular RNA characterization:

- Various circular RNA annotations are publicly available. However, these annotations primarily provide the genomic coordinates of circular RNAs, which often overlap with linear RNA transcripts. As a result, it is challenging to unambiguously assign modifications specifically to circular RNAs.
- Nanopore-based direct detection of circular RNAs remains technically challenging. Current nanopore-based circular RNA detection tools, such as CIRI-long, require special treatment of samples, where we cannot simultaneously identify circular RNA molecules and RNA modifications along them.

Given these constraints, we adopted linear transcriptomic distance in our analysis. However, we fully agree that structural contexts are biologically important. To address this, we have added this limitation in the **Conclusion** part (highlighted) and committed to implementing more robust detection or prediction algorithms in our future analysis.

10. This manuscript analyzed the association between modifications on HepAD38 cell lines. Since MT4 samples are also available, the authors could perform similar analysis on MT4 cell lines as well, and see whether there is cell-line-specific patterns.

Re: We thank the reviewer for this suggestion. In the revised version of the manuscript, we have incorporated the analysis in the MT-4 cell line to explore potential cell-line-specific crosstalk patterns.

Similar to the HepAD38 cell line, we calculated the:

- Distance between modifications
- The modification rate for sites with and without modified neighbors
- The conditional probability for the presence of modifications in adjacent positions.

The finding is consistent with the HepAD38 samples, that:

1. The distance between modifications is significantly shorter than random (**Figure R6a**)
2. The modification rate for sites with modified neighbors is significantly lower than sites without modified neighbors (**Figure R6b**)
3. The majority of pairs of adjacent modification sites, the presence of one modification will reduce the presence of the other modification on the same molecule.

In comparison, in HepAD38 cell lines, we identified 705 negatively and 155 positively associated cases in the binding region of 56 RBPs; in MT-4 cell lines, we identified 489 negatively and 123 positively associated cases in the binding region of 59 RBPs. Detailed information can be found in **Supplementary Table 8** and **Table 9**, respectively. These associated cases are unique to each cell line (no overlap is observed), suggesting a strong cell-line-specific crosstalk pattern.

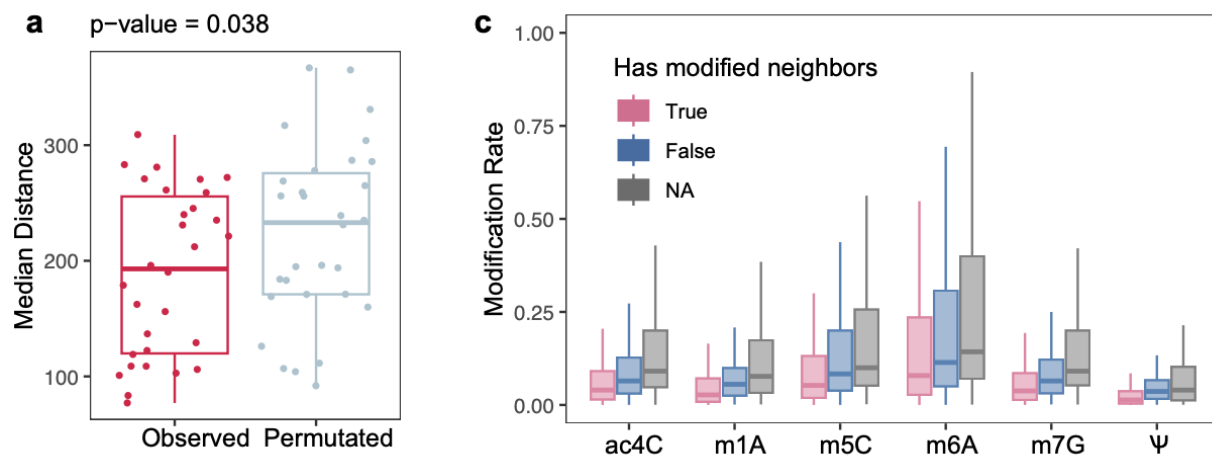


Figure R6. Crosstalk between modifications (a) Transcriptomic distance between two modifications in the MT-4 cell line. **(c)** Modification rate for sites grouped by distance to their nearest modification sites. NA indicates that the site is the only detected modification on that transcript

11. The reliance on NGS-based labels for model training limits applicability to modifications without established profiling methods. Could synthetic RNAs with defined modifications supplement training data?

Re: We sincerely thank the reviewer for this insightful suggestion. We fully acknowledge that the reliance on NGS-based labels limits the detection of modifications without established profiling methods.

However, our current framework prioritizes biological relevance over broad applicability. As illustrated in **Figure R7a** and **Figure R7b**, while synthetic RNAs (e.g., IVT with single modifications) can provide clean signal training data, they lack the natural co-occurrence patterns and cross-talk between modifications observed in native RNAs (e.g., competitive deposition of m6A and m5C). Since our primary goal is to study these biologically meaningful interactions, we preferred to use the gold-standard natural labels for training.

We agree that incorporating IVT data could broaden the applicability. We may integrate both NGS-based labels and IVT data in future work. For example, pre-train the model on IVT data to learn modification signal features, and then fine-tune with biological NGS-labeled data to preserve natural modification patterns.

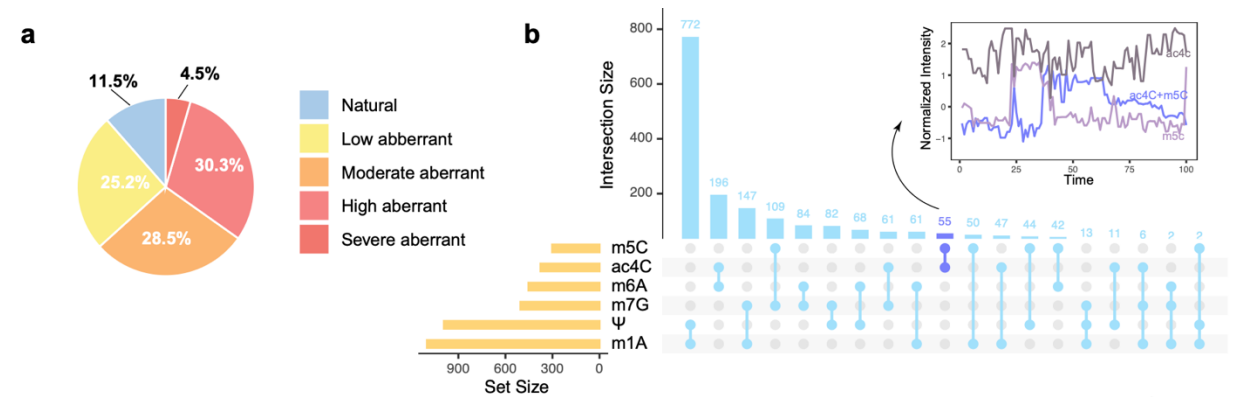


Figure R7. Modification position and identity inference through multi-label model. (a) Pie chart showing the percentage of natural and unnatural IVT 9-mers. natural 9-mers refers to that the modification number is consistent with the NGS observation. Low, moderate, high and severe aberrant 9mers refers to that additional 1, 2, 3-4 and >5 modification number, respectively, from the NGS observation. **(b)** Upset plot showing the co-occurrence of different modifications within 9-mer windows.

12. In terms of the software usage, provides details on the scalability of software. Can DirectRM handle large-scale datasets typical of nanopore sequencing?

Re: Thank you for highlighting this critical aspect of practical usability. DirectRM is explicitly optimized for large-scale direct RNA sequencing datasets through the following computational strategies:

- We take the Pod5 file as input, which requires less storage space and is faster and efficient to query.
- To optimize the memory usage, we split the data into multiple parallel batches and perform downstream analysis as several independent and parallel tasks.
- Optional Pytorch backend for GPU acceleration

We estimated the total time to process 80,000 reads, which takes less than 6 hours to analyze (including preprocessing steps) with a single CPU core and requires less than 1 GB to store relevant files (i.e., input, extracted features, and output). In addition, the capability of DirectRM to handle large-scale datasets was validated on our human HepAD38 and MT4 samples (30-40 million reads per sample).

Below are minor edits and clarifications for the manuscript:

1. The manuscript introduces abbreviations without defining them at first mention, which may hinder readability for non-specialist audiences. Examples include IVET in line 125.

Re: we thank the reviewer for catching this oversight. The term “IVET” has now been formally defined at its first occurrence in line 138 in the revised manuscript.

2. The manuscript occasionally uses different abbreviations for the same metric, which may confuse readers. For example: AUC (Area Under the Curve) and AUROC (Area Under the Receiver Operating Characteristic Curve) are used interchangeably (e.g., line 129 vs. line 124). While both refer to the same concept, consistency is critical for clarity.

Re: We thank the reviewer for highlighting this issue. We have reviewed the manuscript and standardized the use of abbreviations. Specifically, we now consistently use AUROC throughout the manuscript to ensure clarity and avoid confusion.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Re: Thank you for your thoughtful and detailed feedback on this manuscript. I truly appreciate the collaborative effort and the valuable insights provided in your combined report. Your expertise and constructive comments are greatly appreciated and will undoubtedly help improve the quality of the manuscript.

Reviewer #4 (Remarks to the Author):

The paper by Zhang et al studies the crosstalk between RNA modifications with a modified multiple instance learning algorithms, with 3 steps in the module, 1) de novo detection of

modified kmer sites, 2) a binary classification layer for 6 sub-models with each model for one modification type to extract attention-scores for different features, 3) a optimization step that learn integratively multi-labels. Through learning and predicting for multiple modification sites simultaneously, the authors further found that within 100bp, modification level is lower the closer modification site neighbors, and of the modification site neighbor pairs, by comparing marginal probability against conditional probability, the authors showed the positive and negative impact between pairs of modification types, and identified similar results as reported previously. The study is generally interesting and worth reading, however, there are a number of critical concerns regarding the method evaluation and analysis, as well as results presentation. Please see my comments below for more details:

Re: we thank the reviewer for these comments. We have addressed all subsequent comments point-by-point below to further strengthen the manuscript.

Major comments

While the concept of the method is interesting, the overall presentation of DirectRM is confusing and distractive, without clear focus. For example, section 2.1 and 2.2, are less interesting and similar to existing approaches, can be reduced, while 2.4 where comparing independent model and integrative model optimization should be highlighted more. Related to that, the modification position determination results seems unnecessary, suggest to remove this from the manuscript.

Re: We sincerely appreciate your insightful critique regarding the manuscript organization. Your suggestion to shape the methodological focus has significantly improved our presentation. Below, we detail the substantive revision made in response to this comment:

1. We retained **Section 2.1, the de novo modification detection part** in the main manuscript, as de novo detection is a fundamental step of our framework.
2. We moved **Section 2.2: model structure optimization part** to the **Supplementary Note**. The feature importance analysis part was removed in the revised manuscript.
3. We added two new sections that focus on the integrated detection model:
 - **“Modification inference model training with NGS-derived native modification labels”**: this part illustrated the multi-label training process. It comprehensively discussed the label source for model training (i.e., the necessity to use the gold-standard NGS labels) and presented the overall performance of the integrated model.
 - **“Integrated model allows better modification detection than independent models”**: This part discusses the necessity of using an integrated multi-label model rather than an independent binary model.
4. We included the results of benchmark comparison against existing tools in the **“DirectRM: simultaneous identification of multiple RNA modifications”** part.

The evaluation and benchmarking of the method is not convincing. For each module of the method, currently, it seems rather arbitrary, and comparison is within the method itself, hence hard to justify why this is good and why this is better than existing approaches. Suggest to

include other methods, when benchmarking each of the steps, and also when applying to different cell line dataset. Another point is that not many datasets are evaluated here, would suggest adding more datasets in benchmarking, can be public datasets, or new datasets generated, or different species datasets.

Re: We thank the reviewer for this crucial critique regarding methodological benchmarking rigor. We fully agree that comprehensive cross-method and cross-dataset validation is essential, and have substantially expanded our evaluation framework with new datasets. Below, we detail the systematic improvements:

1. We collected six biologically diverse human cell lines: HeLa, HEK293, K562, A549, HepG2, as well as mouse mES samples, from public sources.
2. Regarding the m6A detection, we performed a benchmark against three representative modification detection tools: nanom6A, DENA, and m6Anet. Given the dynamic feature of RNA modifications, for each cell line, we collected cell-specific m6A sites by overlapping the collected single-base m6A sites with cell-specific m6A-MeRIP results (collected from m6A-atlas2). Then, cell-specific m6A sites were used to validate the results by different modification detection tools. Given the varying numbers of m6A sites identified by different tools, we performed 10 rounds of random sampling, selecting 3000 sites from each tool's results in each round. For each round, we calculated the percentage of sites validated by cell-relevant NGS sites.
3. Regarding Ψ detection, due to the limited availability of cell-specific Ψ sites, we adapted to use high-confidence Ψ sites aggregated from available datasets. Similarly, 3000 sites were sampled from the results of each tool and those cell-specific Ψ sites were used to validate the results (repeated for 10 times).

The results of benchmark comparison can be found from the **second paragraph** of the section “**DirectRM: simultaneous identification of multiple RNA modifications**” in the revised manuscript. According to **Table R2**, DirectRM achieved superior accuracy in predicting m6A and Ψ sites across multiple cell lines.

Table R2. Percentage of sites validated by NGS techniques

	Hela	HEK293	k562	A549	HepG2	mES
m6A	DirectRM	39.43%	31.43%	3.95%	48.51%	1.10%
		±0.01%	±0.01%	±0.00%	±0.00%	±0.00%
	m6anet	18.66%	12.71%	1.16%	19.02%	0.06%
		±0.00%	±0.00%	±0.00%	±0.00%	±0.00%
	DENA	12.90%	8.74%	1.09%	13.92%	0.07%
		±0.00%	±0.00%	±0.00%	±0.00%	±0.00%
	nanom6A	23.27%	18.77%	1.78%	24.58%	0.12%
		±0.00%	±0.01%	±0.00%	±0.00%	±0.00%

Ψ	DirectRM	34.04%	22.13%	21.74%	17.48%	15.82%	18.88%
		$\pm 0.01\%$	$\pm 0.01\%$	$\pm 0.01\%$	$\pm 0.01\%$	$\pm 0.01\%$	$\pm 0.00\%$
	nanoMUD	0.05%	0.04%	0.01%	0.01%	0.02%	0.01%
		$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$
	nanoPsu	0.06%	0.03%	0.00%	0.00%	0.00%	0.00%
		$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$	$\pm 0.00\%$

When motivating for the simultaneous modelling of multiple RNA modifications, the description is very unclear, suggest to rephrase this part, to highlight why need to joint model, and why independent model is not working. Currently, the description is limited to remain false positive issues, but false positive issue is a general problem that is with every method, so this is not a very convincing point.

Re: We thank the reviewer for this insightful comment. We have restructured the motivation to explicitly establish why joint modeling is biologically imperative and why independent approaches fail fundamentally:

1. In the “**Modification inference model training with NGS-derived native modification labels**” part, we confirmed the frequent co-occurrence modification events, and they have distinct modification profiles, which necessitates models that account for modification interactions – an integrated multi-label model (**Figure R8a** and **Figure R8b**)
2. In the “**Integrated model allows better modification detection than independent models**” part, we illustrated that the multi-label model achieved better machine learning performance than the independent model (**Figure R8g**).
3. Subsequently, from the site-level evaluation, we grouped the integrated and independent model results based on the number of modified neighbors within 9mer windows. We considered the presence of modified neighbors as a warning of potential false positives. The independent-model result is significantly inconsistent with the gold-standard NGS labels, suggesting its high false-positive rate and indicating that the independent model failed to recapitulate the epitranscriptomic patterns (**Figure R8h**).
4. Moreover, since the independent model was optimized for a single modification type, it may fail to deconvolute overlapped signals. Therefore, the significant inconsistency between NGS patterns and independent-model results is also attributable to this.
5. In contrast, our integrated model, by leveraging the multi-label learning algorithm, can capture the modification dependencies and model those complex signal patterns (e.g., overlapped signals), and make more informed conclusions.
6. We also performed validation on a high-purity IVET dataset, the high error rate of the independent model (20%) versus the 5% error rate of the integrated model also indicates the robustness of the integrated detection model (**Figure R8i**).

Collectively, this evidence demonstrated the necessity of using an integrated multi-label model in our study. We hope the above discussion resolves your concerns.

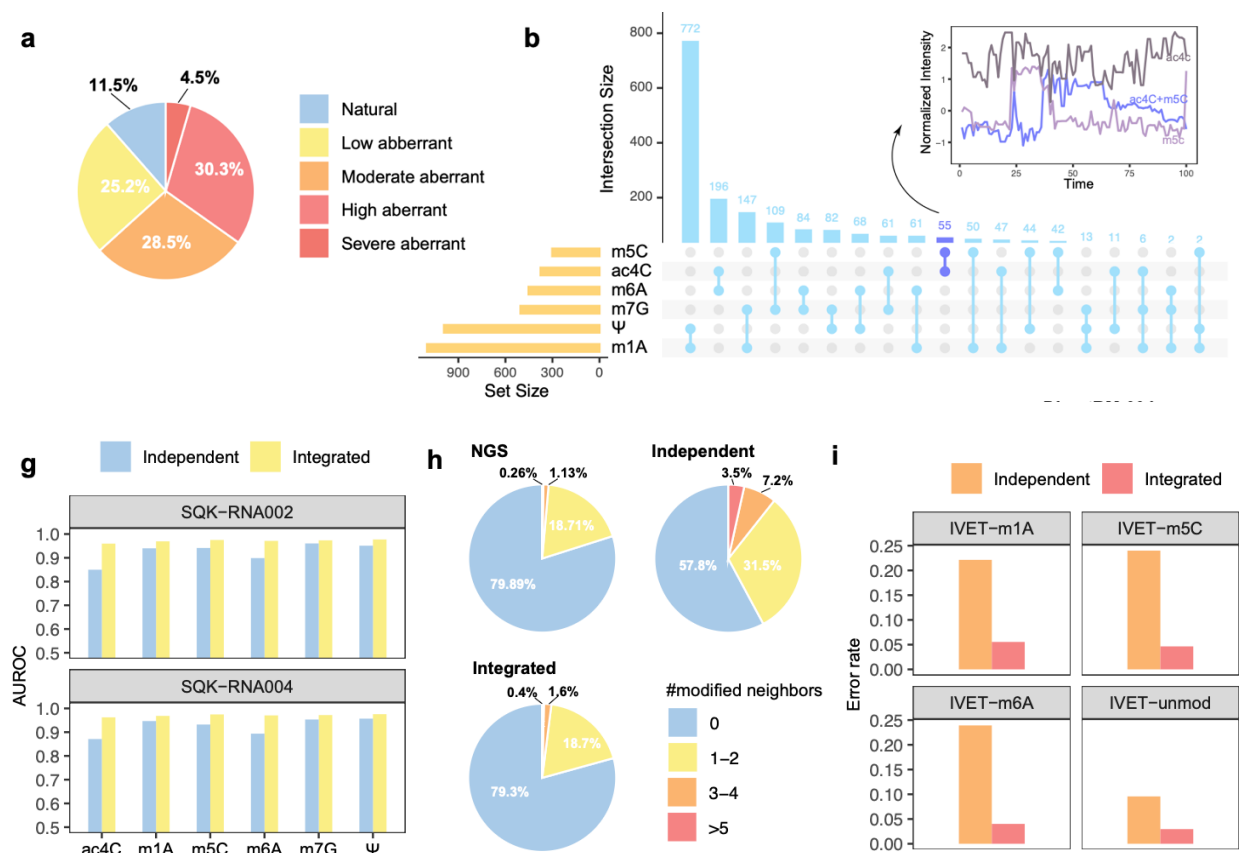


Figure R8. Modification position and identity inference through multi-label model. (a) Pie chart showing the percentage of natural and unnatural IVT 9-mers. natural 9-mers refers to that the modification number is consistent with the NGS observation. Low, moderate, high and severe aberrant 9mers refers to that additional 1, 2, 3-4 and >5 modification number, respectively, from the NGS observation. **(b)** Upset plot showing the co-occurrence of different modifications within 9-mer windows. **(g)** Bar charts showing AUROC scores for independent and integrated multi-label model. **(h)** Pie charts showing the percentage of sites with different number of modified neighbors. **(i)** Bar charts showing the error rate of independent and integrated multi-label under different IVET datasets.

IVT RNA-based training would distort the model because of unnatural residues, also needs more support as to how often IVT RNA would have unnatural residues, if not often, then probably this also reflects the true scenario where consecutive modifications would happen.

Re: We acknowledge the reviewer's valid concern regarding the IVT RNAs.

As highlighted in the “**Modification inference model training with NGS-derived native modification labels**” part in the revised manuscript, we conducted rigorous quantification of unnatural residues in IVT RNAs and the scenarios absent in the IVT dataset. The results establish the necessity and validity of our model training method using gold-standard NGS labels. Our systematic analysis reveals:

1. 90% of residues or 9mers in the standard IVT dataset (complete replacement of nucleotide with modifications) are unnatural (i.e., the modification frequencies exceeding natural observations provided by NGS techniques) (**Figure R8a**).
2. IVT datasets are inherently limited to a single modification type, failing to capture the biologically critical phenomenon of modification co-occurrence. The analysis on the NGS dataset confirmed the frequent modification co-occurrence event and demonstrated that they have distinct modification patterns. IVT dataset is unable to provide training data regarding the co-occurrence of modifications (**Figure R8b**).

Collectively, the systematic analysis established the necessity of using gold-standard natural labels.

Can the authors elaborate how the multi-label learning is done, what are the task and how is this different from single-label learning? What is the main task in the multi-label learning step? The current description in section 4.2.3 is very limited and hard to comprehend.

Re: we thank the reviewer for this crucial critique. We have carefully revised the multi-label training part in the **Materials and Results** section.

Minor comments:

Lines 175-177: The results of comparing the contribution from sequence features against signal and baseball error features are not well analyzed. The conclusion of the sequence feature is more significant seems only applicable to m6A based on Figure 2A and Figure S1. Firstly, there is no test associated with this conclusion. Secondly, based on the figure itself, this is mainly based on sum of each type of features, if all features compared together with a ranking, then it probably can be seen that one or two of the sequence position stands out as very important, with some of the features from signal and baseball error stands out. Would suggest this be reorganised and the results refined.

Re: Thank you for your valuable feedback. After careful consideration, we have decided to remove this section from the main text in the revised manuscript. We agree that the analysis and conclusions in this part were not sufficiently robust and could potentially lead to misinterpretation. By removing this section, we aim to maintain the clarity and focus of the integrated multi-label part.

Line 196: average AUROC>0.9, but Figure 2B all below 0.9, the statement needs to be refined or changed, as it is not reflecting the truth based on the figure.

Re: Thank you for your valuable feedback. We have thoroughly reviewed the entire manuscript to ensure that all statements are consistent with the corresponding figures and tables. Your feedback has been invaluable in helping us improve the accuracy and clarity of our work.

Lines 201-203: The results presented in Figure 2C are confusing. In the first place, why evaluate % of DRACH motif against all m6A sites? Should the absolute number be evaluated for DRACH and total m6A sites separately? Because the % of DRACH is affected by the total

number of identified m6A sites. Even % of DRACH is high, but the number of DRACH is low, then it is not considered good, I think.

Re: Thank you for your thoughtful comment and for highlighting this important point. We fully understand your concern regarding the use of percentages, as the percentage of DRACH motifs relative to all identified m6A sites can be influenced by the total number of identified sites. To address this, we conducted additional analyses to ensure the robustness and fairness of the comparison.

Specifically, for sites identified by each model, we randomly sampled 3000 sites and recalculated the percentage of DRACH motifs. This approach allowed us to control for variations in the total number of identified m6A sites. Importantly, we found that the results obtained from the sampled data were consistent with those derived from the full dataset, confirming the reliability of our conclusions (**Table R3**).

Furthermore, in other instances where percentages were used for comparisons, for example, **the benchmark comparison against the existing tool** part, we adopted the same approach of sampling 3000 sites in addition to using the full dataset. This dual analysis ensures that our findings are not biased by differences in the total number of sites across datasets.

We have updated the manuscript to include these additional analyses and clarified our methodology to address this concern. We hope this resolves your question and provides greater confidence in the robustness of our results.

Table R3. Percentage of DRACH relative to identified m6A sites

	DRACH%	sampled DRACH%
model1	51.41%	52.73%
model2	34.38%	33.80%
model4	31.81%	29.80%
model5	74.53%	74.03%
model8	61.65%	62.70%
NGS	78.11%	77.33%

Line 203-205: results on Figure 2D also need better clarification, why is there a second mode in the end around the end of 3'UTR in all models, but not NGS background profile, this require more discussion and investigation, whether this is related to systematic error in model labels or model specification? Or is it there in the training dataset?

Re: Thank you for your insightful comment. We appreciate your observation regarding the second mode around the end of the 3'UTR in some models, but not in the NGS background profile.

In the process of model comparison in the revised manuscript, we observed that some models exhibit a second mode near the end of the 3'UTR (**Figure R9**). However, the final model we selected (model 5) did not display this issue as prominently, indicating that this phenomenon may be model-specific.

At this stage, we are unable to fully explain the origin of this second mode. It is possible that this phenomenon is related to systematic errors in model labeling, model specification, or other unknown factors. However, through a series of comparisons and validations, we have demonstrated that our final model is reliable and performs well across various metrics. This supports the robustness of our approach despite the observed discrepancies in some models.

We acknowledge that further investigation is required to fully understand the cause of this second mode. Future studies will aim to explore this phenomenon in greater depth, potentially by incorporating additional data or refining the modeling process. We have included this discussion in the revised manuscript (**Last paragraph of “Model structure optimization”** part in the **Supplementary file**) to acknowledge the limitation.

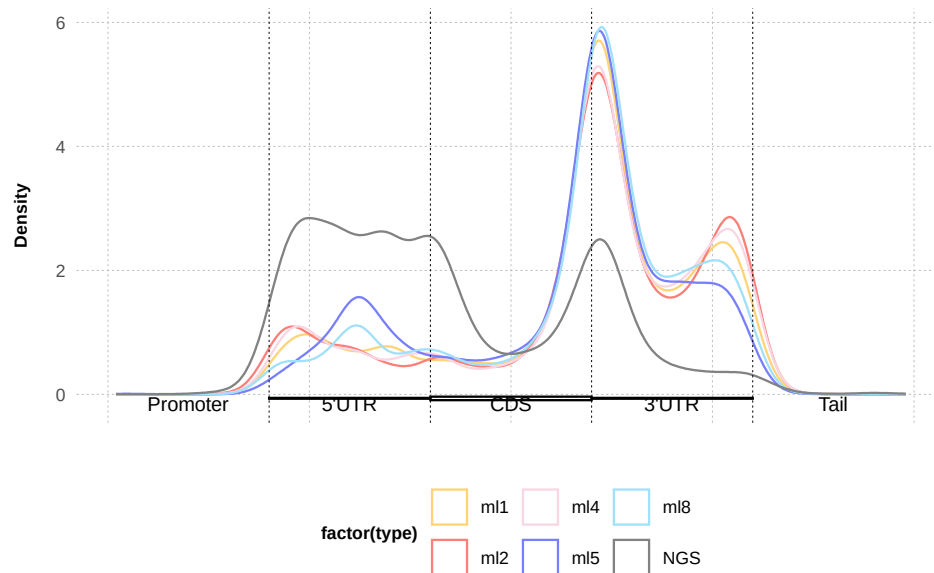


Figure R9. Transcriptomic distributions of identified m6As.

For Figure 3D, the integrated model detects much less significant sites, are they false positives? Could there be true positives missed by the integrated model? This is not evaluated and is worth exploring before a conclusion can be made. Similarly, for lines 247 to 253, this is worth more analysis. Currently, this is all speculation. It would be great if this could be evaluated with

some kind of ground truth label dataset. It probably can be tested with some other synthetic datasets, with a mixture of different degrees of multiple modifications present within one kmer.

Re: Thank you for your thoughtful comment and for highlighting this important aspect of our analysis. We understand your concern regarding the reduced number of significant sites detected by the integrated model and the potential for missed true positives. We also appreciate your suggestion to evaluate the model using high-purity synthetic datasets. Below, we address your concerns in detail:

- While synthetic datasets could indeed provide a controlled environment for evaluation, generating such a dataset (i.e., with a mixture of different degrees of multiple modifications) is highly challenging. We failed to obtain such a dataset at the current stage.
- As an alternative, we use the IVET dataset, which has a single modification type for each dataset. This dataset provides a controlled environment where false positives can be more reliably assessed. Our analysis revealed that the independent model provides more significant sites, its error rate is approximately four times higher than that of our integrated multi-label model (**Figure R8i**). This suggests that the integrated model is more conservative, prioritizing precision over recall, and is likely reducing false positives rather than missing a significant number of true positives.
- We acknowledge that the integrated model's nature may result in some true positives being missed. Based on our analyses, we conclude that the integrated model achieves a better balance between precision and recall compared to independent models, as evidenced by its significantly lower error rate. However, we recognize the need for further evaluation to fully address the possibility of missed true positives. We have added this limitation and future direction to the manuscript, which can be found in the **last paragraph** of the **Conclusion** section.

We have updated the manuscript to include these additional analyses, clarify the rationale for the integrated model's performance, and discuss the limitations and future directions.

Line 80: typo "there still limitations"?

Re: Thank you for pointing this out. We have thoroughly reviewed the entire manuscript to resolve those spelling and grammatical errors.

Line 81: psi, better to use the complete word instead?

Re: Thank you for pointing this out. We have reviewed the entire manuscript to ensure consistency and clarity in terminology. Throughout the text, we now exclusively use the term "pseudouridine" or the specific symbol (Ψ) where appropriate.

Line 81-82: "only report non-canonical bases without specifying their modification types...", need citation specifically to make the statement more clear, currently unclear what the author is trying to say

Re: Thank you for your comment. The intent of this sentence was to highlight that certain ONT-based detection tools report only the presence of modifications (e.g., modified adenosine) but do not provide detailed information about the specific types of modifications (e.g., m6A or m1A).

To address this, we have revised the statement for clarity: “or only report non-canonical bases without specifying their modification types (e.g., modified adenosines)”

Line 87-88: this sentence describing methods being different, and hence hard to collate and combine results, the statement need to be rephrased, maybe the author mean the differences in method makes results highly specific to each method, and hence hard to compare? But what is this statement leading to? Need more clarification

Re: Thank you for your insightful comment. We understand that the original statement may have been unclear and could benefit from rephrasing and additional context to clarify its intent. Below is the revised response to address your concern:

The original sentence aimed to highlight the difference in methodologies across various detection software results in the method-specific outputs, making it challenging to directly compare or integrate results from different studies. To clarify this and provide better context, we have rephrased the sentence as follows:

“In addition, these tools were trained with different groups of features (e.g., base call error features and signal features), algorithms (e.g., supervised and comparative methods), labels, training samples (i.e., in vivo transcribed or in vitro transcribed RNAs) and logics. The differences in detection methodologies lead to results that are highly specific to each method, making it difficult to directly compare or integrate findings across studies. This highlights the need for standardized approaches or integrative frameworks to enable more comprehensive analyses.”

This revised statement better conveys the intended meaning and provides a logical progression to emphasize the importance of addressing these challenges. We hope this clarification resolves your concern, and we appreciate your feedback in helping us improve the manuscript.

Line 152: “coverage.” should be “coverage cutoff.”?

Re: Thank you for pointing this out. After reviewing the context of the sentence, we agree that “coverage cutoff” would be more precise and appropriate in this case. We have updated the text accordingly to ensure clarity and accuracy.

Line 153: with increased coverage cutoff, more modified sites are identified. Because, absolute number of sites with coverage cutoff is reducing, it is the distribution of modified vs unmodified sites become larger. Suggest to rephrase.

Re: Thank you for pointing out the need for clarification. We appreciate the opportunity to further explain our intended meaning and provide a clearer rephrasing.

The original statement aimed to convey that at lower coverage levels, modified reads may not always be detected due to insufficient sequencing depth. However, as coverage increases, the

evaluation of modifications becomes more comprehensive, leading to the identification of more modified sites.

To address this, we have revised the sentence as follows:

"Higher coverage improved recall by enabling more comprehensive detection of modification sites, as insufficient coverage at lower levels may leave modified reads undetected."

Line 207: typo, "is correlate to"

Re: Thank you for pointing this out. We have corrected it to "is correlated to" in the revised manuscript. In addition, we have thoroughly reviewed the entire manuscript to resolve those spelling and grammatical errors.

Figure 3 Legend: missing (E) legend

Re: Thank you for pointing this out. We appreciate your attention to detail. We have carefully reviewed the manuscript to ensure that all legends are complete and accurate.

Line 331: tissues-specific?

Re: Thank you for pointing this out. We have corrected it to "tissue-specific" in the manuscript. In addition, we have thoroughly reviewed the entire manuscript to resolve those spelling and grammatical errors.

Line 336, Figure 4F, should be 5F?

Re: Thank you for pointing this out. We appreciate your attention to detail. We have thoroughly reviewed the manuscript to ensure that all references are consistent and accurate throughout.

Line 338: What does "the dynamic feature RNA modification" mean?

Re: Thank you for your question. We appreciate the opportunity to clarify this point.

By "the dynamic feature of RNA modification," we are referring to the highly dynamic and context-dependent nature of RNA modifications. Specifically, RNA modifications can vary significantly across different cell lines, tissues, or even between biological replications. This variability reflects the intricate regulatory mechanisms that govern RNA modifications and their roles in diverse biological processes.

To provide further clarity, we have revised the sentence to better convey this idea. The updated text are as follows:

"RNA modifications exhibit highly dynamic features, with their presence and patterns varying across different cell lines, tissues, and biological replications, highlighting their context-dependent regulatory roles."

We hope this revision addresses your concern and makes the intended meaning clearer.

Line 373: Tax isoform?

Re: Thank you for pointing this out. The previous version used an expression that was not entirely accurate. "Tax isoform" may have caused confusion, as "Tax" is the name of a gene. To improve clarity and accuracy, we have revised the text to use the term "Tax gene" instead.

Line 397: The sentence is incomplete

Re: We thank the reviewer for pointing this out. We have now revised and completed the sentence to ensure clarity and coherence. The updated sentence now reads: "The color intensity represents the magnitude of the difference between the observed and permuted distances between two modifications."

We apologize for the oversight and have carefully reviewed the manuscript to ensure that all sentences are complete and properly structured. Thank you for bringing this to our attention!

Line 404 Figure 7B: p-value is not provided for the comparison

Re: Thank you for this valuable comment. In the revised version, we have explicitly included the significance level by marking the p-value with "****", which indicates a p-value < 0.0001 (**Figure R10b**). This is a standard convention for representing highly significant differences and is widely used in scientific visualizations. We believe this approach effectively communicates the statistical significance of the comparison while maintaining clarity and visual simplicity in the figure.

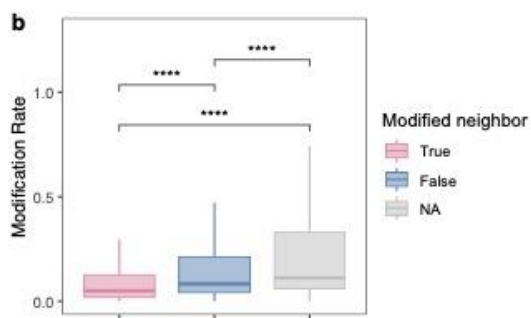


Figure R10. Association between modifications. (b) Modification rates for sites with and without modified neighbors in HepAD38. Sites with other modifications within 100bp region were considered as having modified neighbors. NA indicates that the site was the only site detected on that transcript.

Reviewer #4 (Remarks on code availability):

I have not personally run the code, because of lacking of test dataset. Suggest the authors to include a small test dataset for users to try out.

Thank you for your suggestion to include a small test dataset for users to try out. We appreciate the importance of providing accessible resources for users to validate and explore the code.

To address this, we have included a small test dataset (Extracted nanopore features) named demo_data.zip on our GitHub repository. This dataset is specifically designed to allow users to quickly test and reproduce the key functionalities of the code.

Reference

1. Kovaka, S. et al. Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment. *Nat Methods* **22**, 681-691 (2025).
2. Acera Mateos, P. et al. Prediction of m6A and m5C at single-molecule resolution reveals a transcriptome-wide co-occurrence of RNA modifications. *Nat Commun* **15**, 3899 (2024).
3. Liu, C. et al. IGF2BP3 promotes mRNA degradation through internal m(7)G modification. *Nat Commun* **15**, 7421 (2024).
4. Bellaousov, S., Reuter, J.S., Seetin, M.G. & Mathews, D.H. RNAstructure: Web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res* **41**, W471-474 (2013).
5. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors of DirectRM have improved the manuscript and addressed our previous concerns. The benchmarking between Remora and Uncalled4 revealed that the resquiggling algorithm did not have a substantial impact on the RNA modification classification precision or recall. This flexibility allows others to integrate DirectRM to existing direct RNA nanopore sequencing data processing pipelines.

Re: Thank you for your thoughtful feedback and for acknowledging the improvements made to the manuscript. We are pleased to hear that our revisions have addressed your previous concerns.

Github review:

I do not necessarily think these are required but would be helpful to increase the adoption of DirectRM by the community.

The github page specifies that Dorado v0.6 is required for DirectRM. Is there anything specific to Dorado v0.6 that is required for DirectRM, or is that the only Dorado version that you tested? ONT released Dorado v1.0 this summer, and they usually update their basecalling software once or twice a year. Dorado v1.0 has increased accuracy and read lengths compared to earlier Dorado versions. DirectRM will stay relevant in the coming years if it is compatible with current Dorado versions.

It would be helpful to include a requirements.txt file, a conda environment yml file, a docker or singularity image, or some method to ensure all the required python libraries are properly installed when using DirectRM.

Re: we thank the reviewer for these comments.

Regarding the Dorado version, we apologize for this misleading information. DirectRM was developed and tested with Dorado v0.6, but it would be compatible for other new-released versions. We have now updated our GitHub repository to clarify that while DirectRM was developed and tested with Dorado v0.6, it is also fully compatible with newer versions like Dorado v1.0.

In addition, we provided the Conda environment YML file to ensure the reproducibility of the required environments.

Reviewer #1 (Remarks on code availability):

See above

Reviewer #2 (Remarks to the Author):

I have no further comments.

Re: Thank you for your thoughtful feedback and for acknowledging the improvements made to the manuscript. We are pleased to hear that our revisions have addressed your previous concerns.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Re: Thank you for your thoughtful feedback and for acknowledging the improvements made to the manuscript. We are pleased to hear that our revisions have addressed your previous concerns.

Reviewer #3 (Remarks on code availability):

The code was well organized in github and I successfully installed it and got it to run on the test data set provided.

Re: we thank the reviewer for this comment.

Reviewer #4 (Remarks to the Author):

The authors have satisfactorily addressed all my concerns, and I greatly value their hard work.

Re: Thank you for your thoughtful feedback and for acknowledging the improvements made to the manuscript. We are pleased to hear that our revisions have addressed your previous concerns.