ModiDeC: a multi-RNA modification classifier for direct nanopore sequencing

Nicolò Alagna[1,†], Stefan Mündnich[2,†], Johannes Miedema[1], Stefan Pastore[2], Lioba Lehmann[1], Anna Wierczeiko[1], Johannes Friedrich[1], Lukas Walz[2], Marko Jörg[2], Tamer Butto [2], Kristina Friedland [2], Mark Helm [2,*], Susanne Gerber [1,3,*]

[1]Institute of Human Genetics, University Medical Center Mainz, Mainz 55128, Germany
[2]Institute of Pharmaceutical and Biomedical Science (IPBS), Johannes Gutenberg University Mainz, Mainz 55128, Germany
[3]Institute for Quantitative and Computational Biosciences (IQCB), Mainz 55128, Germany

*To whom correspondence should be addressed. Email: sugerber@uni-mainz.de
Correspondence may also be addressed to Mark Helm. Email: mhelm@uni-mainz.de
†The first two authors should be regarded as Joint First Authors.

Abstract

RNA modifications play a crucial role in various cellular functions. Here, we present ModiDeC, a deep-learning-based classifier able to identify and distinguish multiple RNA modifications (N^6-methyladenosine, inosine, pseudouridine, 2′-O-methylguanosine, and N^1-methyladenosine) using direct RNA sequencing. Alongside ModiDeC, we provide an extensive database of *in vitro*-transcribed and synthetic sequences generated with both the new RNA004 chemistry and the old RNA002 kit. We show that RNA modifications can be accurately recognized and distinguished across different sequence motifs using synthetic data as well as in HEK293T cells and human blood samples. ModiDeC comes with a graphical user interface and an Epi2ME pipeline that allows easy customization and adaptation to specific research questions, such as learning and classifying additional RNA modifications and further sequence motifs. The reproducibility across samples, together with the low rate of false positives, underscores the potential of ModiDeC as a powerful tool for advancing the analysis of the epitranscriptome and RNA modification.

Graphical abstract
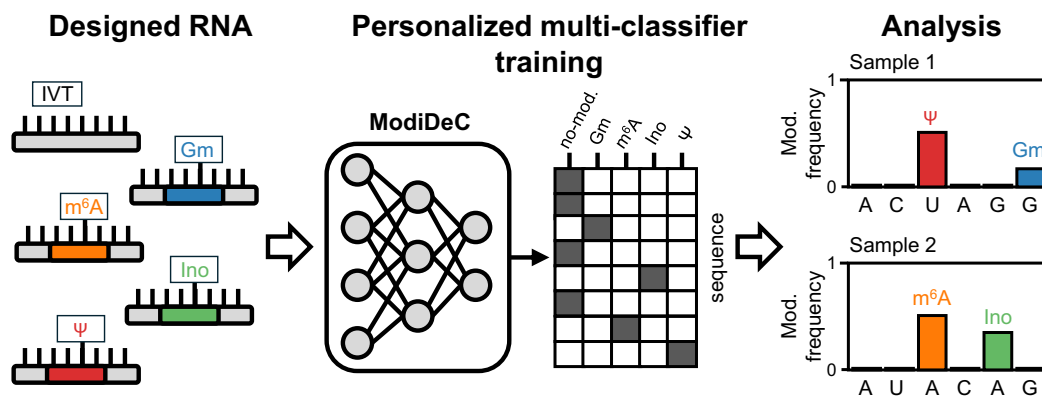
Introduction

Chemical alterations of RNA molecules can affect their structure and stability as well as their interactions with proteins [1–4] and stabilize or alter their metabolism at several stages of the cellular RNA life cycle [5]. Specifically, chemical modification of single nucleotides can interfere with RNA-dependent mechanisms involved in cellular localization, gene transcription, and RNA processing and localization [6], all of which are crucial for protein production and the growth and development of eukaryotic cells [7–11].

More than 170 different chemical modifications of RNA have been identified to date, giving us a glimpse of the complex epitranscriptome landscape [12]. Among the most prominent examples of RNA modifications are pseudouridine (Ψ) [2, 13], inosine (I) [14, 15], 2′-O-methylguanosine (Gm) [4, 16], and the N-methylation at positions 1 and 6 in adenosine (m[1]A [17,

18] and m[6]A [19–22], respectively). For each of the before mentioned modifications, it is known that an abnormal accumulation or an insufficient level of RNA modifications can cause severe cellular dysfunctions and diseases [23–26]. The spectrum of human pathologies triggered by RNA modification disorders ranges from developmental disorders and immunological deficits to neurodegenerative diseases and cancer in its various manifestations [8, 23–26].

Several methods have been developed for investigating the effects of RNA modifications on the transcriptome, including MeRIP-Seq [27], m6ACE-Seq [28], Pseudo-seq [29], miCLIP [30], and GLORI [31]. These methods combine antibodies or chemical treatments with next-generation sequencing for detecting and characterizing transcriptome-wide RNA modifications in short reads. Despite the significant discoveries made using these methods, short-read sequencing technologies frequently struggle to accurately capture the diversity of RNA modifications [1, 32, 33].

Direct RNA sequencing (DRS) is currently revolutionizing the field of epitranscriptomics research as a technology that enables the identification of RNA modification at single-nucleotide resolution in long reads [34–41]. In nanopore sequencing, the passing of modified bases through the nanopores causes substantial changes in the expected current signal, enabling their identification, as impressively demonstrated in recent years [41–44]. In principle, DRS-based methods for detecting modifications can be categorized as comparative or *de novo* detection. Comparative methods such as xPore [45] and DRUMMER [46] compare negative controls with DRS analysis to detect RNA modifications with good performance. However, their requirement for unmodified control data limits their application. *De novo* methods such as m6Anet [47], nanom6A [39], DENA [48], mAFiA [49], or Penguin [50] are based on training personalized deep neural networks using labeled datasets from sequences synthesized *in vitro* or RNAs transcribed *in vivo* to obtain ground-truth labels for modifications. These methods achieve single-base resolution for the analysis of a selected RNA modification, thereby demonstrating a new way to reliably predict RNA modification. In this respect, DRS holds great promise for opening up previously unknown areas of human disease diagnostics and biomarker development, and for bringing a technology-driven revolution to the field of epitranscriptomics.

However, despite the great progress, wider use of the technology has been hampered by several methodological obstacles, particularly the still significant inaccuracy of predictions in biological data, the lack of large-scale gold standard datasets for validation, and the evolving sequencing chemistry. The latter is especially pertinent, as it renders the prediction tools that were specifically trained on certain features of one chemistry completely unsuitable. As a result, analyzing the epitranscriptomic landscape for different types of RNA modifications with DRS remains a challenge [51]. Several recent studies have, for example, highlighted the importance of being able to simultaneously detect multiple RNA modifications [52, 53] and the growing potential of DRS for investigating diverse RNA species and their modifications, including transfer RNA [54], ribosomal RNA (rRNA) [55], and other noncoding RNAs [56]. One example is a pioneering study focused on the interplay between m[6]A and Ψ in messenger RNA (mRNA) [57]. Using a specialized DRS pipeline, they were able to observe both the different effects of these two RNA modifica-

tions in total and polysome-associated mRNA and a change in translation efficiency related to the RNA modifications.

These distinctions between several modifications and the possibility to optimize the prediction accuracy for specific regions or specific RNA types are precisely the technological advances needed if we are to fundamentally understand how a complex multitude of RNA modifications regulate health and disease. In particular, clinical RNA diagnostics requires reliable and easy-to-use options for the detection of RNA modifications that are integrated into standard software.

To overcome this challenge, we developed the Modification Detector Tool (ModiDeC). This flexible and customizable deep neural network can detect, classify, and quantify different modifications. ModiDeC should be considered less as a competitor to other modification base callers that are designed to perform transcriptome-wide scans but more as a customizable tool that can be trained to be a highly specific classifier for certain regions, motifs, and modifications of interest. To offer the possibility of personalized training for individual research questions, we developed a graphical user interface (GUI) and an Epi2ME pipeline that allow the user to retrain ModiDeC to search for user-defined specific sequences or motifs without having to directly interact with the source code.

We demonstrate the performance of ModiDeC in analyzing synthetic sequences and selective RNA pseudouridylation mediated by synthetic designer organelles [58] in HEK293T cell lines. In addition to synthetic approaches, ModiDeC was applied to rRNA data obtained from HEK293T cells and we verified its performance in detecting RNA modifications in human blood samples. Finally, we demonstrate the extendibility of ModiDeC by adding m[1]A and mA (2′-O-methyl-A) to the pool of classifiable modifications. The exemplified fields of application in this study reveal the potential of ModiDeC as a personalizable high-accuracy multi-RNA modification classifier, which can be extended and optimized as required for specific questions.

Materials and methods

Plasmid preparation and *in vitro* transcription

Plasmid sequences were cloned in a pUC57 vector containing an internal T7 promotor, followed by the desired template sequence and a restriction enzyme cutting site (either MscI or SacII) at the end. The linearization reaction was performed overnight following the respective restriction enzyme manufacturer's instructions (Thermo Fisher Scientific). Linear plasmids were purified by phenol–chloroform extraction followed by ethanol precipitation. Linearization of the plasmids and their overall quality were confirmed by separation on an agarose gel and quantified using a spectrophotometer (NanoDrop One).

In vitro transcription (IVT) was performed using a HiScribe T7 High Yield RNA Synthesis Kit according to the manufacturer's instructions. If the transcription product required 5′ phosphorylation for ligation, the concentration of the terminal NTP (GTP, ATP, CTP, or UTP) was reduced, and the reaction was supplemented with NMP (GMP, AMP, CMP, or UMP) at a ratio of 1:5 NTP/NMP.

Transcription products were purified using a Monarch RNA Cleanup Kit (New England Biolabs, T2040), and their quality was assessed by separation on agarose gel and NanoDrop One analysis.

Splinted ligation of RNA constructs

RNA oligos were 5′-phosphorylated to utilize them in the ligation reaction. Phosphorylation was performed using T4 polynucleotide kinase (New England Biolabs, M0201) following manufacturer's instructions. Phosphorylated oligos were purified using Oligo Clean & Concentrator Kit (Zymo Research, D4060) according to the manufacturer's instructions.

Ligations were performed using T4 RNA ligase 2 (New England Biolabs, M0239), combining equimolar amounts of *in vitro*-transcribed RNA and the ligating oligo, 1× T4 RNA ligase buffer, 10% (w/v) PEG-8000, and 10 U T4 RNA ligase 2, as well as a complementary DNA (cDNA) splint ensuring the correct order of ligation. Here, 2% less cDNA splint was supplied compared to the amount of *in vitro*-transcribed RNA or oligos present.

The reaction mixture was prepared as above, omitting T4 RNA ligase 2 for an initial denaturation step at 75°C. After cooling to 25°C over 15 min, T4 RNA ligase 2 was introduced and the reaction was incubated at 16°C overnight.

The ligation reaction was stopped by digestion of the cDNA splint using DNase I (Thermo Fisher Scientific, EN0525) according to the manufacturer's instructions, followed by purification using an Oligo Clean & Concentrator Kit.

The purified ligated constructs were polyadenylated using *Escherichia coli* poly(A) polymerase (New England Biolabs, M0276) according to the manufacturer's instructions, and purified again using an Oligo Clean & Concentrator Kit before proceeding to library preparation. Purity and concentration after each purification step were assessed using a NanoDrop One spectrophotometer.

Isolation of nuclear and cytoplasmic rRNA

Isolation of nuclear and cytoplasmic rRNA was performed as described in Pastore *et al.* 2025 [59]. Briefly, cells were trypsinized, washed with cold phosphate-buffered saline, and resuspended in nuclei isolation buffer [NIB; 10 mM Tris–HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl$_2$, 0.1% IGEPAL CA-630, 0.1% Tween 20, 1% BSA (bovine serum albumin), 0.15 mM spermine, 0.15 mM spermidine, and 0.2 U/µl RNase inhibitor]. Homogenization was performed with a loose pestle (10 strokes), followed by incubation for 15 min on ice and then centrifugation at 1190 × *g* for 5 min at 4°C. The cytoplasmic fraction (supernatant) was collected, and Trizol was added at room temperature (RT). The remaining pellet was washed twice with 500 µl of ice-cold NIB, centrifuged, and resuspended in 300 µl of ice-cold NIB. For isolation of nuclei, 300 µl of 50% Optiprep solution (Stemcell Technologies, 07820) was mixed with the homogenate to form a 25% Optiprep mix. This was layered over 30% and 40% Optiprep gradients in a 2-ml tube, centrifuged at 20800 × *g* for 20 min at 4°C, and nuclei (∼600 µl) were collected from the 30%–40% interface. The nuclei were washed with nuclei wash buffer (10 mM Tris–HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl$_2$, 0.1% Tween 20, 1% BSA, and 0.2 U/µl RNase inhibitor), centrifuged at 1635 × *g* for 10 min at 4°C, and the supernatant was removed. Nuclease-free water was added to the nuclei to a volume of 500 µl and 500 µl of Trizol was added prior to RNA isolation.

For RNA isolation, nuclear and cytoplasmic fractions (∼1 ml) were incubated at RT for at least 5 min. Chloroform (200 µl) was added to the nuclear and cytoplasmic fractions. Samples were vortexed for 15 s, incubated at RT for 3 min, and centrifuged for 15 min at 20800 × *g* at 4°C. The upper aqueous phase (∼550 µl) was transferred into a fresh Eppi and 500 µl of isopropanol was added, thoroughly resuspended, and incubated at RT for 15 min. Samples were centrifuged for 10 min at 20800 × *g* at 4°C. The supernatant was discarded, and the pellet was washed with 75% ethanol in nuclease-free water, followed by centrifugation for 5 min at 7500 × *g* at 4°C. The supernatant was discarded, and the pellet was air dried for 5–10 min. The RNA pellet was dissolved in 30 µl of RNase-free water and mixed in a high-shear rotor-stator sample mixer (Hula Mixer, Thermo Fisher Scientific) for 10 min at RT, and RNA concentration was quantified using the Qubit dsDNA Assay Kit (Invitrogen, Q32851).

DRS library preparation

Libraries were prepared with 1 µg of polyadenylated input material using either SQK-RNA002 or SQK-RNA004 kits (Oxford Nanopore Technologies) according to the manufacturer's standardized protocol. The prepared libraries were quantified using the Qubit dsDNA Assay Kit (Invitrogen, Q32851). Libraries were loaded onto MinION flow cells and sequenced with MinKNOW (version 24.02.26) set to fast live base calling until no more sequencing pores were available.

Peripheral blood

Peripheral blood was obtained from a healthy volunteer. RNA was extracted using a PAXgene Blood miRNA Kit (Qiagen) according to the manufacturer's protocol, except that the RNA was eluted in nuclease-free water instead of the provided buffer. The RNA was characterized using the Bioanalyzer (Agilent) total RNA nano assay according to the manufacturer's protocol. The RNA had a concentration of 363 ng/µl and an RNA integrity number of 7.2. Globin mRNA was depleted with the GLOBINclear-Human Kit (Thermo Fisher Scientific, AM1980) according to the manufacturer's protocol; this was performed four times on a total input amount of 20 µg of RNA to yield 11 µg of globin-mRNA-depleted RNA. The concentration was measured using a Qubit RNA HS Assay (Thermo Fisher Scientific). Two micrograms of RNA were stored for later use in the direct RNA run. The remaining 9 µg of RNA was split into three aliquots of 3 µg, respectively, and used for poly(A) selection with the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs) according to the manufacturer's protocol. The poly(A) selection resulted in a total yield of 23 ng, as measured with the aforementioned Bioanalyzer assay. The sample had an average length of ∼1 kb. The sample concentration was measured again using the aforementioned Qubit assay. Reverse transcription, polymerase chain reaction (PCR), IVT, polyadenylation, and 5′ capping were carried out according to the methods of Tavakoli *et al.* [43] with modifications as follows. Each IVT primer was used in the PCR at a final concentration of 0.5 µM. The input amount of mRNA used for reverse transcription and PCR was 7.1 ng. The output amount of cDNA was 905 ng, as measured with a Qubit DNA HS Assay (Thermo Fisher Scientific). IVT was performed twice, each with an input of 126.7 ng of cDNA. The outputs were pooled to yield 4.9 µg of RNA, as measured with the Qubit RNA HS Assay. Libraries were prepared using the direct RNA sequencing Kit (SQK-RNA004, Oxford Nanopore Technologies). The library output was 167 ng of RNA/cDNA hybrid (Qubit DNA HS Assay). The complete library was loaded onto the PromethION RNA Flow Cell (FLO-PRO004RA).

Neural network training and data processing

ModiDeC is a two-input neural network with ~20M parameters (see Supplementary Fig. S13) that combines personalized and new structured inception blocks with long short-term memory (LSTM) to classify RNA modifications. To train the neural network, raw signals and the respective one-hot encoded sequences are necessary. A minimal number of passages is required for the neural network to obtain the correct training outcome. The raw pod5 files from all sequencing runs were first base-called using the high-accuracy model in Dorado v0.7.3 using the flag –emit-moves. Next, the raw signal was preprocessed by re-squiggling to the reference the Z-score normalized raw signal using Remora (Oxford Nanopore Technologies). The re-squiggled signal was then split in chunks containing 400 raw-signal measurement points and the corresponding sequence converted in one-hot encoding representation was assigned to each of them. Signal chunks and the one-hot encoded sequences (maximum sequence length of 40 nt) were used for the first (input 1) and second (input 2) inputs of ModiDeC, respectively. No more preprocessing steps were used for the input generation. In parallel, the one-hot encoding output of the neural network is also generated, in which each base is converted into its respective modified variant or into an unmodified label. A mixture of modified and unmodified data was essential for training ModiDeC, including IVT data for training the neural network. Ultimately, the neural network was trained for three epochs with a learning rate of 0.0001 using the Adam optimizer.

To generate the training data for ModiDeC, the "ModiDeC data curation" GUI was used to preprocess 30000 reads for each motif used in this work (Fig. 1E). Ten thousand high-quality IVT reads (selected using samtools and filter flag $q = 20$) that aligned to the gen-code v46 transcripts were also preprocessed using the GUI and then used for the training. For the sample analysis, we established a minimum threshold of 500 reads for each transcript to ensure statistical reliability. This value was decreased to 80 for IVT data as the number of reads produced during IVT measurement can be small.

Results

Creating a training dataset for ModiDeC

To train a neural network to identify multiple modifications at different sequence motifs and distinguish them from one another, the neural network must be presented with various modifications in different sequence contexts. In this training dataset, the sequence, location, and modification type must be precisely known for each construct to form the so-called "ground truth." To provide such training data, we designed and sequenced chemically synthesized oligonucleotides, allowing us to pick and choose the modified nucleotide as well as the sequence context. By ligating those oligonucleotides into a larger construct suitable for DRS, which performs optimally with long reads, we created multiple point-modified sequences with complete information regarding their modification status. Details on the design of the training data are provided in the Supplementary Section S1. Ten oligonucleotides were created per modification type (m6A, Gm, I, and Ψ; see Supplementary Table S1) to give a total of 40 different oligonucleotides. Each oligonucleotide contained a 9-mer motif flanking the modified nucleoside, along with the same unmodified 9-mer sequence,

such that the training included both modified and unmodified motifs.

For oligonucleotides containing m6A, the sequences were based on the evolutionarily conserved DRACH motif [19, 60]. To ensure that ModiDeC distinguishes modifications based on their specific signal rather than the surrounding sequence context, one DRACH-motif-containing sequence was selected to represent a single oligonucleotide motif across all modification types. For the remaining motifs, we either mirrored known modification sites in the human transcriptome, as reported in the Modomics database [14], or chose naturally occurring sequences but without a known modification background. A list of all sequence motifs is provided in Supplementary Table S1. The oligonucleotides were then incorporated between two *in vitro*-transcribed RNA sequences via splinted ligation, creating a larger sequence construct ready for sequencing (see the "Materials and methods" section and Supplementary Fig. S1B). Following this strategy, for all 40 modified constructs, we also created a corresponding negative control without any modification.

The RNA004 flow cell was introduced by Oxford Nanopore Technologies for use on their DRS platform in 2023. Its new chemistry differs significantly from the SQK-RNA002 kit in several ways [61], notably in *k*-mer length, dwell time, current level, and translocation speed. Because ModiDeC is intended for use with data acquired using the old RNA002 or the new RNA004 sequencing chemistry, we double-sequenced all 40 constructs (once with the SQK-RNA004 kit and once with SQK-RNA002), together with the unmodified control, thus creating two identical datasets that differed only with respect to the kits/flow cells used in DRS. This allowed us to train an accurate model for each chemistry, giving the user the flexibility to analyze any DRS datasets with ModiDeC, regardless of the chemistry used (Fig. 1A).

In addition, we designed five constructs for the initial testing of ModiDeC. These constructs contained either no modification, one in a previously unknown motif, or two distinct modifications in motifs the model had been exposed to in training. These sequences were also generated by splinted ligation, and in the case of doubly modified constructs, the final construct sequence alternated between the oligonucleotides and the IVT sequences to ensure sufficient distance between the modified oligos. For details of the ligation process, see the "Materials and methods" section and Supplementary Fig. S1C.

Training and validation of ModiDeC neural network

ModiDeC is a personalized two-input neural network that combines LSTM layers with a newly designed Inception-ResNet block to classify multiple RNA modifications simultaneously on the data. In detail, the neural network remaps the sequence by analyzing the signal, and each sequenced nucleotide is classified as either "unmodified," Ψ, Gm, I, or m6A. ModiDeC was trained using the 40 splint-ligated sequences, each containing one modified oligonucleotide (Supplementary Table S1 and Supplementary Fig. S1A and B). Thirty thousand reads per modification and motif were used for training. This number was chosen following a pre-analysis investigating the change in prediction accuracy as a function of altering the amount of training data. As expected, the percentage of correctly accessed modification sites increases with the amount of training data (see Supplementary Section S2 for more details). However, 90%
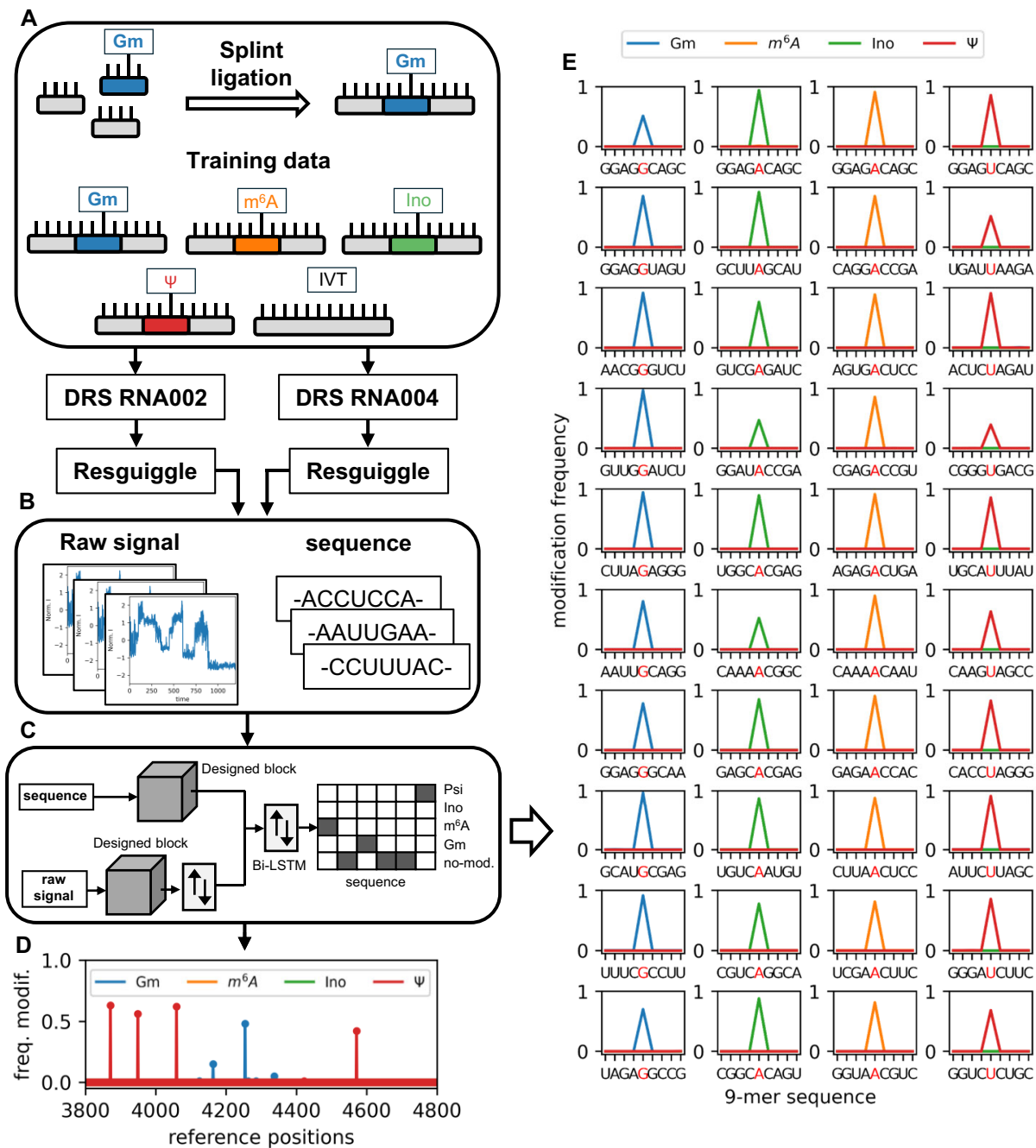
Figure 1. Schematic representation of the workflow—from data generation to preprocessing, model training, and prediction. (**A**) Representation of oligos used for training ModiDeC. Modified and unmodified labeled reads were used for the training. (**B**) After base calling, data are re-squiggled to the reference and chunks with respective references are sent to the neural network. (**C**) Scheme of the ModiDeC neural network. (**D**) Example output after neural network analysis. (**E**) The modification frequency on the test data for the 40 motifs used in this work (RNA004 kit). The red letter is the modified base. Each panel shows the predicted frequency of modification (*y*-axis) with the respective modification type and modified motif reference (*x*-axis) in the oligos. For clarity, we visualize only the core motif of the 9-mer. The red letter indicates the modified base.

accuracy can already be achieved with ∼5000 reads. With more reads, the accuracy increases only minimally and almost reaches a plateau at 30000 reads.

After base calling, each read was re-squiggled to the reference, cut into chunks of 400 time points, linked to their sequence, and then used for the input of the neural network (Fig. 1B; see also the "Materials and methods" section for more details). For each chunk, an output label for the neural network was created in parallel. The output of the neural network is a one-hot encoding representation, where each base is classified between "unmodified" or one of the four modification types (m⁶A, Gm, Ino, and Ψ). In total, >40 million labeled chunks were used to train ModiDeC (Fig. 1C). The final accuracy of the neural network to identify the correct modification position after the training phase was higher than 91%. An example of a ModiDeC analysis output is shown in Fig. 1D. Af-

ter selecting a pod5 file, the corresponding aligned bam file, the reference, and the number of reads to analyze, ModiDeC remaps each read and finally reports the frequency of predicted modifications at the single-nucleotide level for each position in the reference.

We validated ModiDeC on a new dataset that contained all the motifs and modifications used during training but had never been seen or analyzed by the neural network. Five thousand reads were used for each oligo. All test data were sequenced with SQK-RNA004. Figure 1E shows the results of the validation analysis. The network correctly identified the position and type of modification 100% of the time for the 40 sequence–modification combinations. The rate of correct quantification of the occurring modification (detection of the modification frequency) was 81%. However, this average prediction accuracy is both motif- and modification-dependent. Although we can report impressive prediction accuracies of almost 100% for some motif–modification combinations, others were obviously more difficult to detect. This might also be partly due to the sequencing runs differing slightly in quality. In each individual case, however, a modification was clearly identified (even if correct quantification was not always achieved). A detailed overview of true-positive, false-negative, and modification miscalling predictions for each motif–modification combination is reported in Supplementary Table S2.

Particularly interesting in this context are the results for the DRACH motif AGACA, in which a different modification was introduced at the same position in each case (Fig. 1E, first row). The data show that ModiDeC correctly predicts that the A base in this DRACH motif is modified. However, because we have inserted different modifications at the same motif, we can exclude that ModiDeC has learned only from the sequence in the training data that the DRACH motif, for example, stands for m^{6}A modifications. Instead, we can show that ModiDeC can correctly distinguish and classify between several modifications, even if a new modification appears on an actual "typical" motif for another modification. Because ModiDeC makes an independent prediction for each individual read at the single-molecule level, even a mixed prediction could also have been made. However, ModiDeC was able to distinguish precisely between the four modifications in every case. To determine the false-positive prediction rate precisely, we validated ModiDeC on the negative control data. We observed that ModiDeC has an extremely low false-positive prediction rate between 0% and 1% (Supplementary Fig. S3).

Next, we wanted to test if similar results could be reproduced in the dataset sequenced using SQK-RNA002. To this end, we retrained and validated the neural network using the same datasets but sequenced with the old chemistry. Also, with SQK-RNA002-acquired data, ModiDeC shows great ability to predict precisely the position and the corresponding modification for each individual motif–modification combination (for details of this analysis, see Supplementary Figs S4 and S5, and Supplementary Table S3). Interestingly, the SQK-RNA002 dataset also shows that some of the combinations can be excellently predicted, but for others, the network tends to more frequently misclassify modifications as "unmodified." However, these are not the same combinations as in SQK-RNA004, which again suggests a difference in the quality of the sequencing performance rather than a systematic bias in the network's predictions. Regarding the determination of the false-positive prediction rate, a similar result emerges from an analysis of the unmodified control data sequenced with SQK-RNA002 to that we observed with SQK-RNA004. Although we can discern a slight difference between the performances of the models trained with the old and new chemistries, the amount of false-positive predictions is still low, ranging between 0% and 5.5%.

In conclusion, the new RNA004 chemistry gives a slight improvement in detection accuracy, but ModiDeC works reliably with both the old and new chemistries. Of particular note here is the high accuracy of the position prediction (base-level resolution), as well as almost perfect classification. Since no major differences were found between the two kits in terms of ModiDeC's performance, and the SQK-RNA002 kit is no longer manufactured, we decided to focus on the results of the SQK-RNA004 kit for further analysis in this study.

Next, we tested ModiDeC on five RNA constructs previously unknown to the network and with varying complexity of modification. These constructs either carried no modifications at all (Fig. 2A), carried one modification (Fig. 2B and C), or contained two adjacent modifications at a specific position in the sequence (Fig. 2D and E). These five constructs were designed as a random sequence, in which a total of six of the motifs examined here (unmodified or modified) were inserted into the sequences at specific positions (Fig. 2). For the unmodified reads, the neural network correctly returned zero modified positions (Fig. 2A). For the singly modified reads, ModiDeC identified constructs composed of two oligos, one of which contained a position (73) modified as Ψ (Fig. 2B and C). Oligos used for the construction were identical in sequence between constructs and therefore only differed by which oligonucleotide contained the modified position. ModiDeC was able to correctly identify the positions modified and classify perfectly the type of modification linked to those positions with an accuracy higher than 86%. The fact that both constructs shared identical oligonucleotide sequences highlights that ModiDeC accurately detects the state of modification rather than assigns them based on sequence patterns. Next, the systems of greatest complexity—those doubly modified—were analyzed using ModiDeC (Fig. 2D and E). For both oligos, the neural network was again able to identify correctly the positions of modification and classify the modification type linked to those positions with high accuracy (>85% of all modifications could be correctly detected on a single-nucleotide level).

From the test analysis, we observed that ModiDeC always detected the modification position with a high frequency of modification. However, it seems to underestimate the expected value by ~10% on average. We decided to investigate the limit of detection (LOD) and limit of quantification (LOQ) of ModiDeC by virtual titration analysis (see Supplementary Fig. S2G–J). The underlying idea here is to generate test batches with selected modification frequencies by mixing modified and unmodified reads. We generated 11 titration batches for each of the four modification types (from 0% to 100%, with a variation of 10%, expected modification frequency) and analyzed the data using ModiDeC to see whether the predicted and expected modification frequencies matched. For titration batches of a respective modification type, modified and unmodified reads were randomly selected to generate a batch size of 1000 reads from the test data (see Fig. 1E and Supplementary Fig. S3 for test data). After this, the LOD and LOQ were calculated following the formula (see Equations 1 and 2 in Supplementary data) after a linear regression
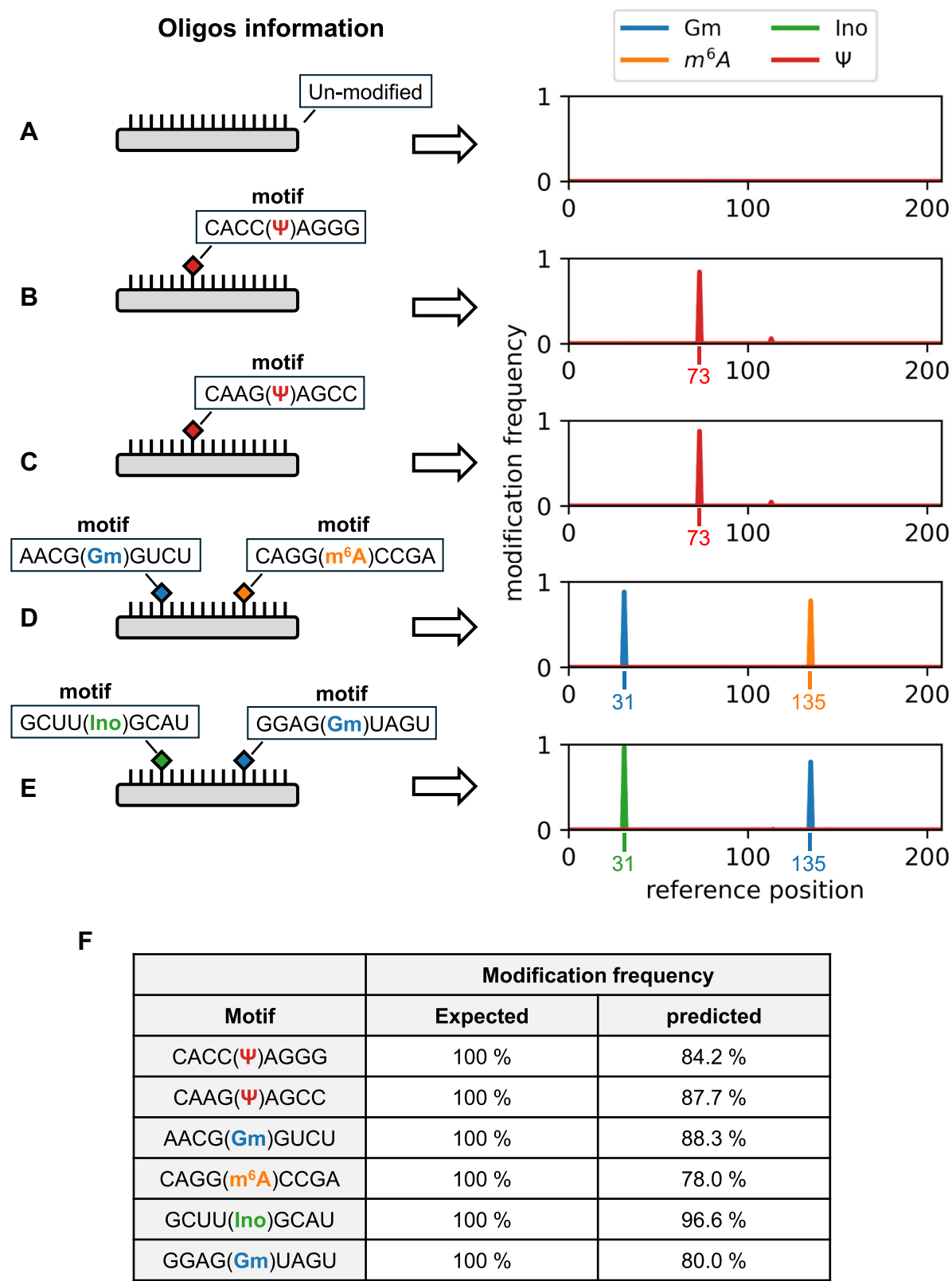
Figure 2. ModiDeC analysis on new RNA oligos for reads with varying complexity. (**A**) unmodified, (**B**, **C**) singly modified, and (**D**, **E**) doubly modified reads. (**F**) Table showing the expected and ModiDeC-predicted modification frequency for each motif.

Motif	Modification frequency	
	Expected	predicted
CACC(Ψ)AGGG	100 %	84.2 %
CAAG(Ψ)AGCC	100 %	87.7 %
AACG(Gm)GUCU	100 %	88.3 %
CAGG(m⁶A)CCGA	100 %	78.0 %
GCUU(Ino)GCAU	100 %	96.6 %
GGAG(Gm)UAGU	100 %	80.0 %

fit. The analysis shows similar LOD and LOQ values for the four RNA modification types, with average values of $LOD_{avg.}$ = 1.1% and $LOQ_{avg.}$ = 3.2%.

Next, we compared ModiDeC with Remora. In detail, we trained two Remora models, one for Gm and one for inosine modification type, using the same training data as for ModiDeC (see Supplementary Section S2) and compared the two neural network results on the test oligos containing an inosine and a Gm modification (Fig. 2E). The results show that ModiDeC could obtain the expected position with no additional false positive, while Remora shows several false positives over the construct (Supplementary Fig. S2K–M). The latter reveals a higher efficiency on training data and a lower false-positive rate (FPR) of ModiDeC compared to Remora for the given dataset.

HEK293T cell analysis: selective organelle and rRNA

To test the ability of ModiDeC to detect and quantify modifications in physiological datasets, we continued our study using data obtained from HEK293T cells. The first example is an already characterized "optimized RNA-editing organelles (OREO)" [58] that in combination with guide RNAs can selectively pseudouridylate a target RNA. The OREO system has been tested on a selectivity reporter harboring two target positions in mCherry and EGFP fluorescence protein sequences, in which the former is preferentially modified based on an optimized recruiting system of the organelle. In detail, the organelles preferentially select the EGFP RNA, block pseudouridinylation in a specific position in the sequence, and reduce the amount of fluorescent protein produced by the cell. This means that by measuring the amount of pseudouridine in specific target sites, it is possible to quantify and compare the amount of EGFP and mCherry in the cell. In this regard, the selectivity of pseudouridinylation was confirmed by measuring the amount of pseudouridine in specific target sites of mCherry (position 565) and EGFP (position 115) mRNAs in HEK293T samples transfected either with the designer organelle (sample O) or without the organelle (sample nO) using DRS together with the pseudouridine detection model (Oxford Nanopore Technologies).

ModiDeC was used to reanalyze the EGFP and mCherry datasets to track and quantify pseudouridine in three of the published samples, namely the sample with the organelle (O), without the organelle (nO), and a control sample without any targeting system (C) for both RNA target sites in the selectivity reporter. For this purpose, ModiDeC was previously exposed to the pseudouridine target positions modified in both the training and testing phases (see Figs 1E and 2B and C). The results are shown in Fig. 3A. We found that ModiDeC correctly predicted the expected position (red numbers in Fig. 3A) and assigned a Ψ modification to the U nucleotide. Comparing the nO sample with the O sample, we observed a reduction in Ψ from 36% to 13.8% (2.61-fold) at the EGFP site, whereas for mCherry the reduction was from 73% to 56% (1.3-fold). The control sample also showed Ψ peaks for EGFP (∼11%) and mCherry (∼3%). These values are consistent with those published and experimentally validated in the original work by Schartel *et al.* in 2024 [56].

Next, we calculated the mCherry/EGFP pseudouridine ratio, a measure of organelle selectivity, for both OREO-containing and OREO-free systems for Ψ counts predicted by ModiDeC and compared it to the ratio derived from the Ψ-detection model implemented in Oxford Nanopore Technologies' most recent base-caller Dorado (Fig. 3B). Although that model was able to detect the Ψ sites, it did not reproduce the expected trend. In contrast, the predicted mCherry/EGFP pseuduridylation ratio was lower for the OREO-containing system (12.7-fold) compared to the OREO-free sample (16.9-fold; see Supplementary Fig. S6 for Dorado-Ψ-caller values). This unexpected result can be mainly explained by the fact that the Ψ model from Dorado exclusively considers U-called sites and therefore overlooks mismatched bases. As shown by Schartel *et al.* and others, Ψ can cause the motif-specific base-calling error of a U–C mismatch in DRS data [43, 58]. Only by manually counting the U–C mismatches and adding these values to the results of the Dorado pseudouridine caller, the expected increase in the mCherry/EGFP ratio from 2.3-fold to 6.7-fold was observed upon switching from the OREO-free to the OREO-containing system (see Supplementary Fig. S6 for Ψ and U–C mismatch Dorado values).

In contrast, ModiDeC was able to reproduce the increase in mCherry/EGFP pseudouridinylation ratio in sample O versus sample nO, by considering the detection of Ψ only. We also calculated the mCherry/EGFP ratio from a ModiDeC analysis performed in the same way as in the original study [58], where the Ψ values obtained from the nontargeting control system were subtracted (baseline correction) from the pseudouridine counts of both organelle and no-organelle systems (Fig. 3C). The baseline-corrected ModiDeC analysis showed fold changes of 2.8 and 17.6 for no-organelle and organelle systems, respectively, which is consistent with the validating bisulfite-induced deletion sequencing values reported [58].

Next, we analyzed a fragment of 18S rRNA (Fig. 4A; for a complete structure of 18S, see Supplementary Fig. S7). We chose this region because it contains documented modification sites [62]. In our synthetic training dataset (Fig. 1D), we integrated four motifs for Ψ and three motifs for Gm, which cover known and isolated modification sites on 18S rRNA (see table in Fig. 4B). We analyzed two *in vitro* samples (sample 1, rRNA from the cytoplasm; sample 2, rRNA from the nucleus) of the 18S derived from HEK293 cells together with synthetic 18S IVT reads as a negative control (Fig. 4C). Because both samples originate from the same cell line, albeit from different compartments, we expect a conservation of modification sites at varying frequency of modification. The ModiDeC analysis of samples 1 and 2 shows a highly congruent prediction of Ψ- and Gm-modified sites, precisely matching the trained motifs (Fig. 4C, upper and middle panels). The prediction of modification frequency showed the expected variations between comparable loci of different conditions (e.g. peaks 295 and 927; see Fig. 4B), while maintaining similar dynamics between distinct loci within a sample. The classification of Ψ- and Gm-modified sites was confirmed for every predicted site. In addition, the analysis of 18S IVT data (Fig. 4C, bottom graph) shows a drastic reduction of the modification frequency at modification sites matching trained motifs, indicating a low FPR. Although rRNA has several modified sites [62], ModiDeC was trained only on a small number of rRNA motifs containing a modification (see table in Fig. 4), which were correctly detected during the analysis. This result, together with the low FPR shown in the IVT analysis, indicates the high selectivity of the neural network. Next, we wanted to compare the Ψ modification frequency detected by MoDiDeC
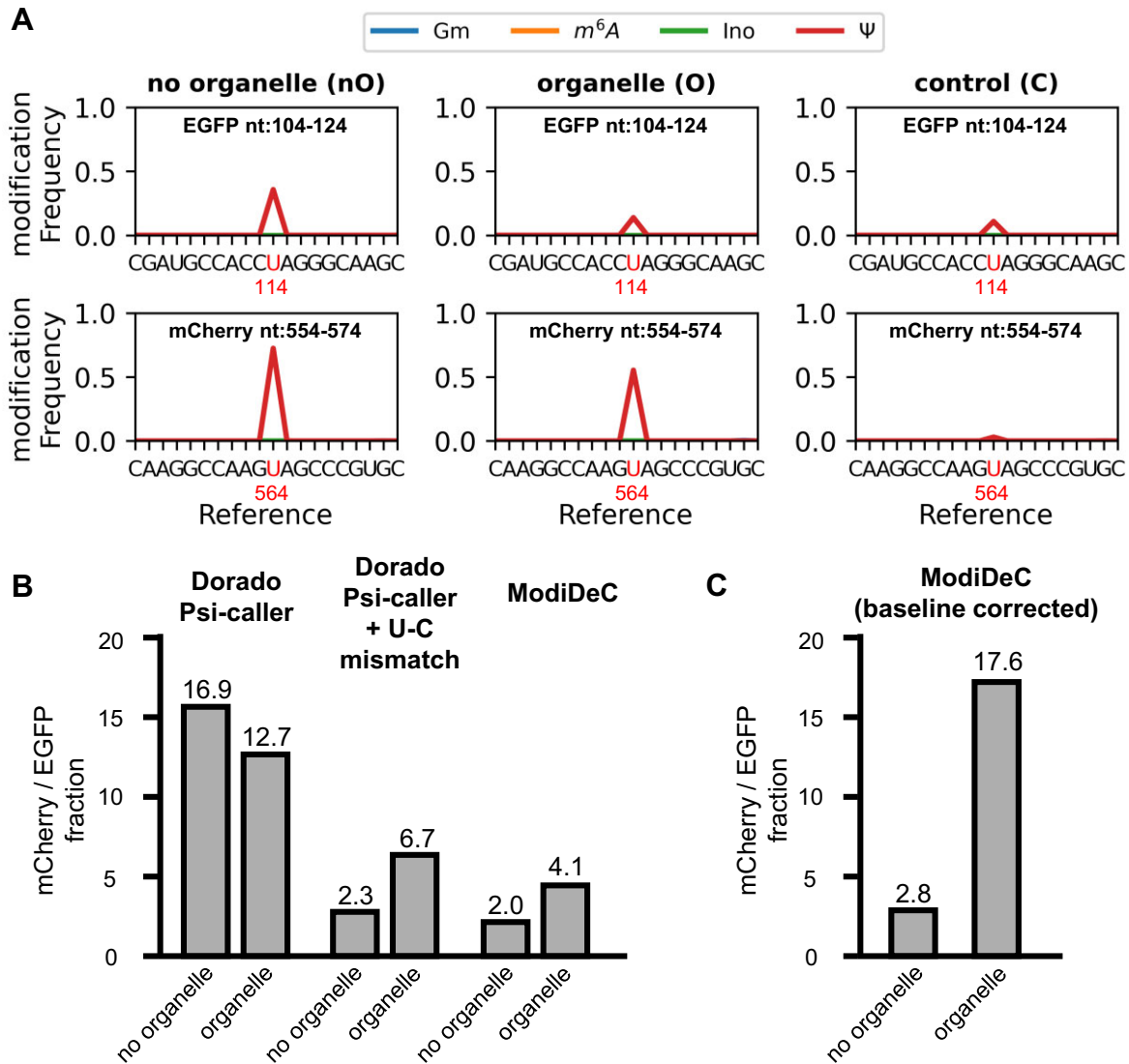
Figure 3. HEK293T OREO system analysis using ModiDeC and Dorado. (**A**) ModiDeC analysis on HEK293T RNA samples without organelle (left), with organelle (center) and control (right) for EGFP (top row) and mCherry (bottom row). Red numbers and letters indicate the expected modified base and position. (**B**) mCherry/EGFP ratio calculated for organelle and no-organelle systems using Dorado pseudouridine model (Dorado-Ψ-caller), Dorado-Ψ-caller + U–C mismatch, and ModiDeC. Expected trends can be tracked with Dorado-Ψ-caller + U–C mismatch and ModiDeC. (**C**) mCherry/EGFP ratio from ModiDeC after subtracting the control (baseline correction) for organelle and no-organelle systems.

with both mass spectrometry (MS) measurements [62] and the Dorado Ψ model (Supplementary Table S4) to understand how well ModiDeC quantifies target modifications. The comparison shows that ModiDeC and Dorado show similar results for these sites, while both exhibit values that deviate to varying degrees from the expected MS reference values, either approaching or diverging from them, depending on the target position values. These results suggest that ModiDeC can quantify, with good approximation, modified sites in *in vitro* samples.

In summary, the analysis of the physiological HEK293T cell data in the two examples above validates the performance and customizability of ModiDeC. Once trained on a specific context, it can be used for the quantification of modifications in biological systems (Fig. 3) as well as for precise detection and classification of position (Fig. 4). ModiDeC also showed impressive reproducibility in its results.

Detecting RNA modification in human blood

Next, we were interested in the general applicability of ModiDeC for diagnostic purposes. To this end, we tested ModiDeC on human transcriptome data from the most common sample type in routine diagnostics, i.e. peripheral blood. It is important to mention that ModiDeC should be considered less as a competitor to other modification base callers that are designed to perform transcriptome-wide scans but more as a customizable tool that can be trained to be a highly specific classifier for certain regions, motifs, and modifications of interest.

The following strategy was chosen to assess the accuracy of the forecasts: We analyzed and compared both human peripheral blood from a healthy subject and the generated whole-transcriptome-wide IVT data from the same individual as a negative control. ModiDeC predictions were then validated with available experimental data from the literature.
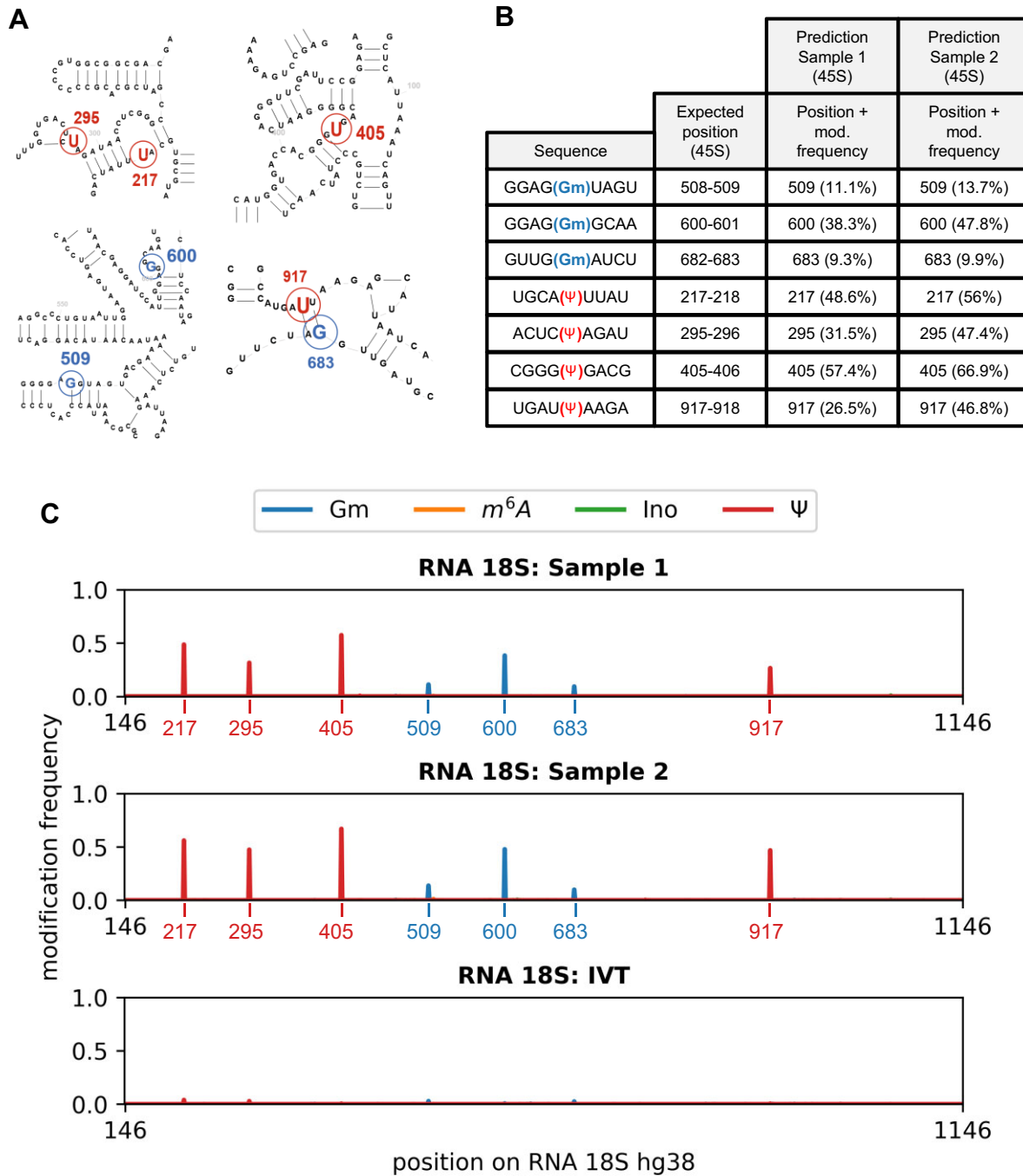
A

B

Sequence	Expected position (45S)	Prediction Sample 1 (45S) Position + mod. frequency	Prediction Sample 2 (45S) Position + mod. frequency
GGAG(**Gm**)UAGU	508-509	509 (11.1%)	509 (13.7%)
GGAG(**Gm**)GCAA	600-601	600 (38.3%)	600 (47.8%)
GUUG(**Gm**)AUCU	682-683	683 (9.3%)	683 (9.9%)
UGCA(**Ψ**)UUAU	217-218	217 (48.6%)	217 (56%)
ACUC(**Ψ**)AGAU	295-296	295 (31.5%)	295 (47.4%)
CGGG(**Ψ**)GACG	405-406	405 (57.4%)	405 (66.9%)
UGAU(**Ψ**)AAGA	917-918	917 (26.5%)	917 (46.8%)

C

Figure 4. Analysis of 18S rRNA. (**A**) RNA45SN1 RNA structure. The highlighted letters are the modified bases recognized by ModiDeC (red = Ψ, blue = Gm). The numbers indicate the reference position. (**B**) Table showing expected motif sequence, expected position (45S), and the ModiDeC analysis on samples 1 and 2. ModiDeC correctly predicts the positions, showing significant values for modification frequencies. (**C**) ModiDeC predictions for RNA34SN1 hg38 for sample 1 (top), sample 2 (middle), and IVT (bottom). Both samples 1 and 2 show distinct Ψ and Gm peaks compared to IVT.

As an example, we chose the well-studied and annotated TP53 tumor suppressor, responsible for multifunctional transcription factors, such as DNA repair or cell cycle arrest, and known to interact with the epitranscriptomic network of m⁶A [63–65]. ModiDeC was used to analyze the 33 transcripts linked to the TP53 gene with the purpose of identifying m⁶A RNA modification sites that are indexed in the RMBase v3.0 database [66]. It should be noted that the annotations of the RMBase v3.0 have not been experimentally

validated for peripheral blood, but only for human cell lines (HEK293). Although a tissue-specific methylation landscape must be assumed, we have nevertheless used this resource for validation. Since no database for m⁶A sites in human peripheral blood currently exists, this resource is the one that most closely matches our data in the literature.

The database reports 49 possible m⁶A RNA positions on those transcripts linked to TP53. Aligning these positions with ModiDeC m⁶A training motifs, the analysis could down-

stream to 12 reported m^6A RNA positions that have a strong overlap with the training motifs, which we have therefore considered in the following. These 12 possible m^6A sites were correctly detected by ModiDeC, and m^6A peaks were predicted at four of these sites in three different transcripts (Fig. 5A–E). The dashed black lines in Fig. 5B–E indicate the expected position of the m^6A site as reported in the RMBase v3.0 database. All four positions have an exact overlap with the expected position. Moreover, the negative controls rule out a false positive at the same positions in the IVT data. In addition to the four annotated m^6A sites, ModiDeC has detected three further m^6A positions (Supplementary Fig. S8) that might be very interesting candidates for further experimental validations. The other 37 possible m^6A RNA positions annotated in the RMBase v3.0 database were also investigated using ModiDeC. The analysis reported no additional m^6A positions detected at these sites. However, as mentioned above, these sites show significant differences in motif sequence compared to those used for training ModiDeC (overlap between 20% and 30%), underlining the high selectivity of the neural network on trained motifs.

To show that ModiDeC is also suitable for predicting the existence of all four modifications (m^6A, Gm, Ψ, and I) simultaneously in our human blood sample, we also analyzed a second group of transcripts, together with the matching IVT data (see Supplementary Figs S9 and S10). Here again, we were particularly interested in transcripts that are potentially clinically relevant. For example, the transcripts shown in Supplementary Fig. S9A–E are derived from genes known to be dysregulated in cancer. For example, CD147 (Supplementary Fig. S9A) plays an important role in the development of various malignant tumors [66], while RAB11B (Supplementary Fig. S9E) is a member of the RAS oncogene family [65]. Regarding potential validation, we were able to validate the m^6A sites with the help of the m^6A-Atlas in the due course [67]. According to the table in Supplementary Fig. S9F, all the predicted m^6A motifs were confirmed by the reference database; most of the predicted positions match the annotated ones exactly. Only two of the predicted positions are less precise, having positional shifts of 1 and 6 nt. However, we cannot confirm whether this observed deviation is due to ModiDeC making an inaccurate shift in the prediction or whether the annotation is imprecise at this point due to inexact experimental resolution.

To the best of our knowledge, there are currently no data resources for the other predictions that we could have used to reliably confirm the validations for inosine and Gm. Therefore, we refer to these only in the Supplementary data and note that, from our point of view, they are highly interesting points for further targeted experimental validation steps. Following this strategy, we can therefore envision ModiDeC as a tool for the targeted prediction of previously unexplored modifications.

Another interesting finding is that ModiDeC can also identify positions where the sequence differs slightly from the trained motif (Supplementary Table S5). All listed positions contain at least one nucleotide that differs from the initially trained motifs (as listed in the Supplementary Section S1). Some positions even contain partial combinations of the training sequences. This suggests that ModiDeC recognizes features of modifications in sequences that are even beyond the set of motifs used for training.

Next, we wanted to show that ModiDeC can also recognize further modifications, such as m^1A sites. For validating the predictions in the biological data, we selected a specific N^1-methylation of adenosine (m^1A) on the mitochondrial *ND5 RNA*, which is described in the literature for its biomedical relevance [18]. Due to the methodological difficulties in using commercially available nonspecific antibodies against m^1A, the sites of m^1A RNA methylation predicted to date are highly controversial. Using a more stringent transcriptome-wide mapping approach, at single-base resolution, m^1A was only detected in a low number of cytosolic mRNAs [17, 68]. Irrespective of the approach used, numerous studies have clearly identified mitochondrial ND5 RNA modified with m^1A at position 1374 [17]. However, the current mapping approach is rather complicated and time consuming. The ability to identify m^1A sites based on DRS would be groundbreaking for the field.

Therefore, we expanded the pool of classifiable RNA modifications by introducing two additional motifs into the ModiDeC analysis: the first was a randomly chosen natural motif without a known modification context (CCACm̲^{1} ̲AAACC), while the second was the specific motif of the mitochondrial ND5 RNA transcript sequence (UGGCm̲^{1} ̲AGCCG) [18] (Fig. 5F). The full sequences, and that of the unmodified control, are shown in Supplementary Table S1. The predictions on the validation data (generated using the same splint-ligation strategy as before), together with the frequency of detection of these motifs in the evaluation data, are shown in Fig. 5G. For our modified validation dataset, the average frequency recognition was ∼82% and the modified site was accurately detected in both cases. We also detected the m^1A site in a new sample (having a sequence previously unknown to the network) that contains the motif with an m^1A modification at position 70 (Fig. 5H, top) and compared it with the results from the unmodified control sample (Fig. 5H, bottom). ModiDeC correctly detected the position of m^1A with a frequency of around 71% (Fig. 5H, top). In the control sample (Fig. 5H, bottom), almost every nucleotide is correctly classified as "unmodified" except a small peak with a frequency below 5% at position 76, which might indicate a threshold for the reliability of results from an m^1A analysis.

After the successful training with the modified oligos, the m^1A-expanded version of ModiDeC was then used to analyze the mitochondrial RNA that served as a template for the training motif. For this purpose, we again used the blood of the healthy subject and specifically targeted the mitochondrial ND5 RNA transcript (Fig. 5I). ModiDeC analysis showed an m^1A peak with a modification frequency around 11% at the expected position of 1374. The IVT data of the same transcript were analyzed, and an m^1A prediction of <2% was found in the same position; we considered this a false positive.

Expanding ModiDeC to m^1A modification detection, we could observe that the neural network can classify correctly RNA modifications with similar chemical structure (m^1A and m^6A). To further investigate the capability to distinguish between different modifications of the same ribonucleotide, ModiDeC was trained to detect m^1A, m^6A, and 2′-*O*-methyl-A (mA), which are three RNA modifications having similar chemical structures. The three modifications were inserted in the same sequence context, where the same 9-mer (UGGCAGCCG) contains one of the modifications in
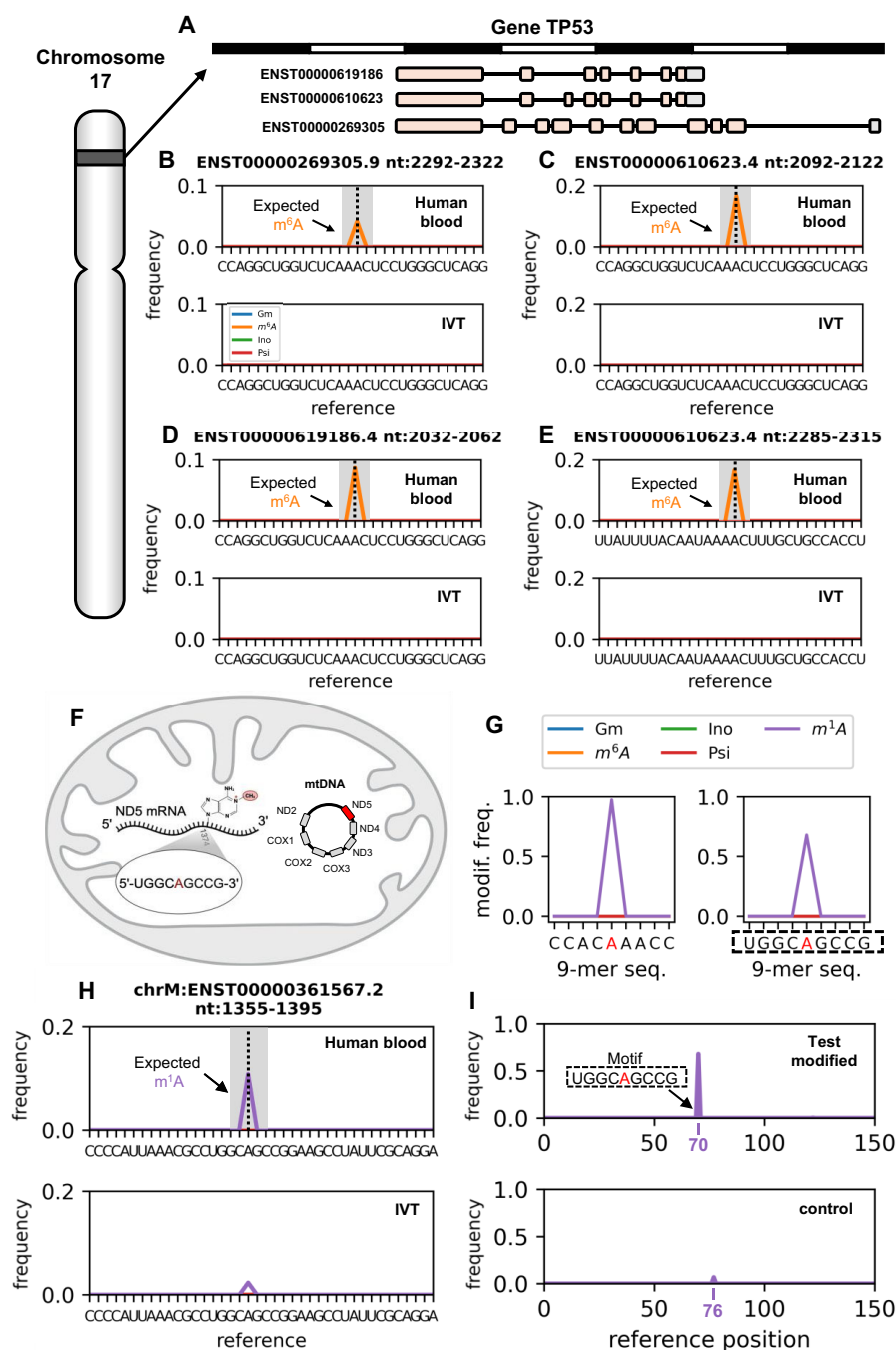
Figure 5. Detection of site-specific RNA modification in human peripheral blood samples. (**A**) TP53 gene cartoon representation from chromosome 17. (**B**–**E**) ModiDeC analysis on four different transcripts linked to the TP53 gene for the human blood sample and IVT datasets. While modification peaks are detected at the expected positions in the human blood, predictions on the IVT show very low values similar to a baseline. (**F**) Mitochondrial ND5 RNA showing the motif containing the m^{1}A modification (red "A"). (**G**) ModiDeC validation on two different 9-mers (9-mers are shown in the *x*-axis). The red letter indicates the modified base in the validation data. (**H**) ModiDeC results on modified test data (top) and control sample (bottom). The peak at position 70 indicates a modification classified as m^{1}A. The sequence is the mitochondrial RNA motif, with the modified base indicated in red. (**I**) ModiDeC analysis on a mitochondrial human blood sample (top) and mitochondrial IVT data (bottom). ModiDeC can detect the expected m^{1}A peak associated with the mitochondrial ND5 RNA.

the middle. ModiDeC analysis on test oligos shows that the neural network can classify correctly the three modification types, which all show high modification frequency as expected (Supplementary Fig. S11). Moreover, the analysis on unmodified oligos shows a low false-positive detection of ∼4%.

The generation of false positives by ModiDeC was also investigated by analyzing at random 200 IVT transcripts (each with >80 reads per transcript) from the chromosomes here studied. In detail, we analyzed two things: (i) the maximum FPR for each transcript (Supplementary Fig. S12A) and (ii) the average amount of false positive detected per transcript and their value range (Supplementary Fig. S12B). The FPR analysis shows that ∼99% of the data have a maximum false-positive value below 6%, and ∼90% below 4%, while only 1% have a false-positive value between 6% and 8%. We also calculated the mean value $\mu = 1.9\%$ and standard deviation $\sigma = 1.5\%$, suggesting a ModiDeC minimum threshold ≈5% for biological data (calculated by 2σ from the μ value). The analysis shows that ModiDeC has on average 200 false positive-predicted sites for a transcript length of 2191 nt. Moreover, the analysis shows that most of the false positive-predicted sites show a modification frequency value between 1% and 4%.

In Summary, on the human data ModiDeC was able to make precise and reproducible predictions for m^{1}A and m^{6}A for mitochondrial and TP53 (chromosome 17) transcripts, respectively. Additionally, using synthetic oligos and human blood data, we demonstrated that ModiDeC can be easily extended with further modification classes (m^{1}A and mA) and motifs in addition to the initial 40 sequences on which it was originally pretrained (see Supplementary Table S1).

ModiDeC GUI: a GUI for retraining the model

To enable users, regardless of their expertise in bioinformatics, to use and expand ModiDeC, we created a user-friendly pipeline that includes three GUIs (Fig. 6). These GUIs facilitate the use of the model for analyzing specific locations or RNA types to personalize it, or to retrain it without having to directly interact with the source code. The pipeline consists of three parts: "ModiDeC data curation," "ModiDeC training," and "ModiDeC analysis." "ModiDeC data curation" (Fig. 6, top) supports the user in generating a dataset for re-training the neural network. This GUI is highly customizable and allows the user to preprocess data by linking output labels to signal information for the proper generation of training data. The "ModiDeC training" GUI (Fig. 6, middle) allows the selection of neural network parameters, such as batch size and epochs, and to retrain the ModiDeC for the specific use. The "ModiDeC analysis" GUI can be used to visualize modification sites and corresponding modification frequencies on the reference sequence (Fig. 6, bottom). A more detailed explanation of the GUI inputs can be found in Supplementary Section S8. Additionally, a tutorial for using the three GUIs can be found on the GitHub page (see the "Data availability" section). In parallel, we also developed an Epi2ME "ModiDeC pipelines" that allows the user to retrain ModiDeC also in the ONT platform. Similar to the GUI, you can find three pipelines of ModiDeC called as the three GUIs ("ModiDeC data curation," "ModiDeC training," and "ModiDeC analysis"). The three pipelines follow the same steps of the GUIs, but allow the user to adapt ModiDeC to a specific problem using the Epi2ME program from ONT.

Discussion

The neural network ModiDeC was developed to recognize whether an RNA nucleotide is modified at a specific position and assign it a specific modification. ModiDeC can be used with DRS data sequenced with an SQK-RNA002 sequencing kit, as well as with those acquired with the newer RNA 004 chemistry.

To demonstrate its performance, ModiDeC was initially trained on four different types of RNA modifications (m^{6}A, Gm, I, and Ψ) with 40 different motifs (10 for each type of modification). Expanding this fundament, we demonstrated that ModiDeC can be simply trained for further motifs or new modifications, exemplified for a mitochondrial m^{1}A modification.

In this respect, ModiDeC should not be considered a competitor to approaches that perform global scans for a specific modification but as a highly precise, complementary tool for accurately determining and validating certain combinations of motif and modification, with a biological and/or clinical context. Its strength is its ability to determine the exact position and type of a modification, especially if there is great uncertainty about whether a modification even exists in a particular region. Using validation data with known ground-truth and unmodified IVT data, we showed that ModiDeC could correctly identify the type of modification with 100% accuracy, even if different modifications at different locations were to be detected in the same sample or even within the same motif (see Fig. 1E; we added all four modifications once each at the same position in the DRACH motif "AGACA").

ModiDeC clearly distinguishes itself from other excellent tools already available for detecting RNA modifications that have been trained, for example, with randomer oligonucleotides "randomers" (as e.g. in the case of the Dorado base callers) and are optimized for large-scale screenings for only one particular type of modification. Although methods trained with randomers offer great advantage if the aim is to find as many modifications as possible to obtain an overview of the entire epitranscriptomic landscape, they are less favorable if one is interested in a specific motif and wants to achieve the highest possible accuracy in terms of position and quantification. Randomly generated sequences have the limitation (purely statistically) that some motifs will occur particularly frequently in the training data and will, therefore, be recognized particularly well, while other motifs will be encountered relatively rarely in the training data and will, therefore, be learned particularly poorly. However, by using randomer-based approaches, it is very challenging and demanding to influence which motifs are learned particularly well. It cannot be assumed that exceptionally well-trained motifs are biologically significant; indeed, underrepresented motifs might be precisely those that are of particular interest to the application-oriented user.

As also apparent with our own training data, even under identical conditions, and with a very large amount of training data (30 000 identical sequences per motif–modification combination), some motifs were harder to learn and already showed a high false-negative rate on the synthetic data, whereas other motifs were much easier to remember (see Fig. 1E and Supplementary Fig. S4). If such a "difficult" motif is then also underrepresented in randomized training data, it will most likely have a poor detection rate when the model is applied to biological data.
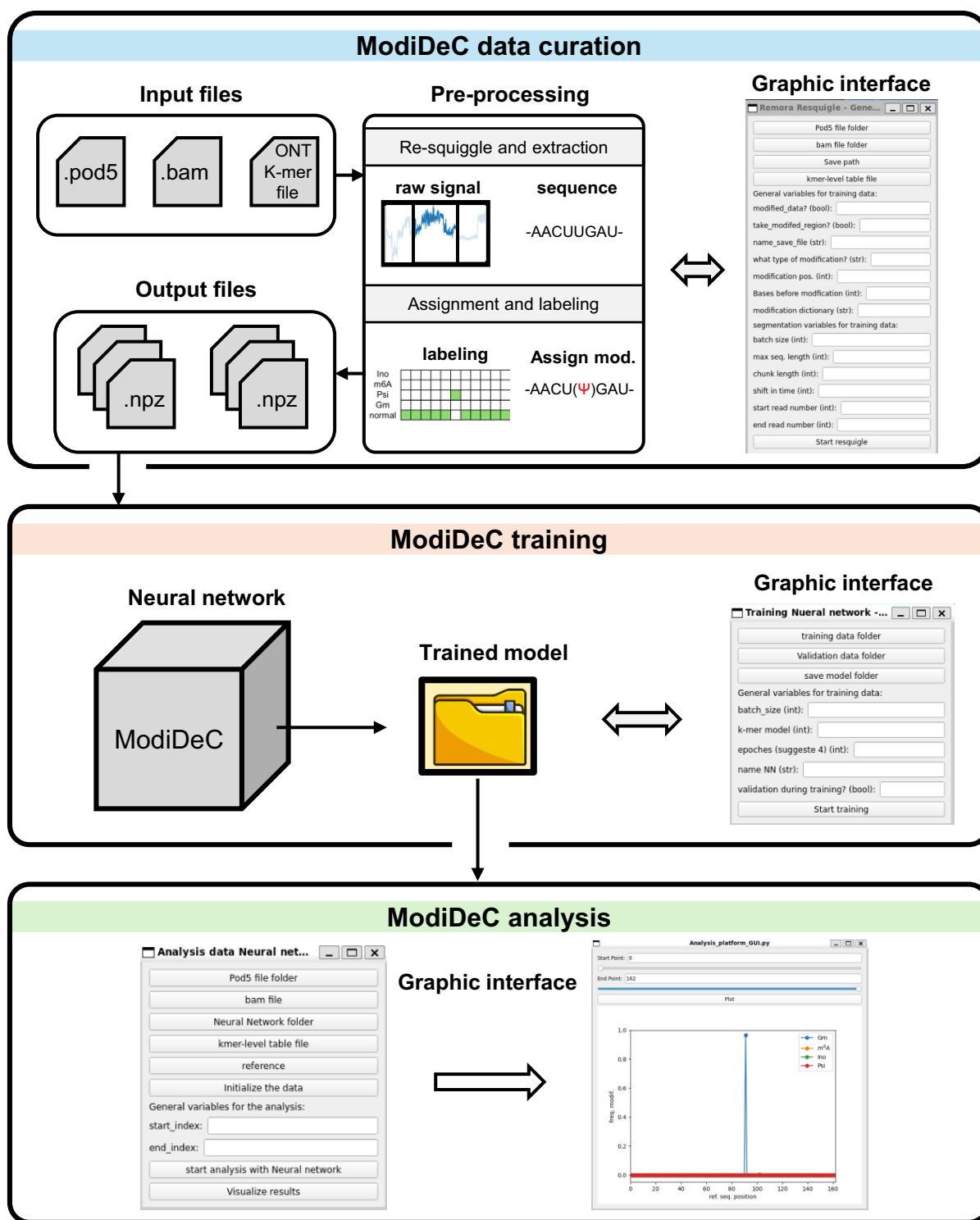
Figure 6. ModiDeC workflow to adapt the neural network to specific biological analysis using ModiDeC's GUI. Schematic representation of the ModiDeC's GUI pipeline. The GUI is divided into three communicative sub-parts, created for retraining the neural network or directly using it for data analysis. (Top) "ModiDeC data curation." Raw data can be preprocessed to correctly create the input and output for the neural network training. (Middle) "ModiDeC training." Helps the user to easily retrain the ModiDeC for their problem. (Bottom) "ModiDeC analysis." Interface created to visualize and store the model analysis.

In a further complication, false-positive predictions, false negatives, and misclassified RNA modifications for transcriptome-wide predictions can rarely be validated owing to a lack of gold-standard datasets. For m⁶A, larger experimentally validated benchmarking datasets for HEK293T cell lines already exist [31, 52, 67], which allows predictions about modifications to be validated to a certain extent. Also, a reference catalog has been recently published for pseudouridine [43, 62]. However, these experimentally validated databases should not be regarded as gold standards either, as false-positive or false-negative predictions cannot be ruled out, as exemplified by several contradictory statements regarding the existence of certain modifications. Even for the already very well-familiar m⁶A landscape of HEK293T cell lines, considerable uncertainty exists, as shown by the overlap between the CHEUI base-caller data and the GLORI database, which was tested in two independent studies in 2024 [49, 52]. Chan *et al.* reported a site-level overlap of 0.64 and Mateos *et al.* determined it to be 0.85. Similar observations have been made for pseudouridine sites [69] and were also currently reported with respect to m¹A sites [17, 68]. Furthermore, a considerably tissue-specific and even time-dependent modification landscape must be assumed, which makes validation in other tissues or cell types highly speculative.

In this context, a tool such as ModiDeC can be of great utility, as it can be applied explicitly to ill-defined and/or difficult-to-detect modifications and motifs for validation. In cases for which we knew the exact truth and were thus able to validate the predictions (Figs 1–5), ModiDeC correctly recognized the presence of a modification in one nucleotide every time, correctly identified the type of modification each time, and also recognized the exact position of the modification in the validation data, even if the sequence surrounding the motif was completely new because the algorithm had been trained on different training data with a different sequence context. As demonstrated in several of the biological samples, we were able to show that when interested in specific issues or regions, it is not necessary to first generate all possible motifs and permutations around a particular modification. Rather, we have already achieved very significant performance with a very limited number of motifs (10 per modification).

Peak detection was made further reliable by continually training and validating our findings with matching IVT data that had a negligible FPR for the annotated positions (<2%) and a low FPR on the overall IVT transcripts (<5% false positives in >97% of the cases). A result of particular note is that ModiDeC seems to have versatility in identifying motifs that differ from those used for training but contain variations in the sequence (see Supplementary Table S5). This indicates that ModiDeC has potential for identifying features that have not been explicitly learned from the training pool and would thus also be suitable for use in human data, in which a large number of genetic variants, such as single nucleotide polymorphisms, must be assumed. We also see ModiDeC as particularly suitable for predicting unknown modification sites, which can then be experimentally validated. As shown in Fig. 5 and the Supplementary data (Supplementary Fig. S8), instances arose where ModiDeC predicted Gm or inosine that could not be validated due to a lack of reference data in the literature, but which due to the experimental setup (high predictions frequency for the blood samples, with simultaneous correctly recognized negative IVT) are interesting candidates for further studies.

A major unresolved issue arises from systematically comparing models that predict modifications from DRS data for use in addressing biological questions. Most of the models already developed have been trained on RNA002 chemistry and only a few on data from RNA004 sequencing. Therefore, only a few benchmarks are representative, as these require that models are compared on differently generated data (some on RNA002 and some on RNA04). Additionally, all would have the problem that an exact validation, apart from synthetic oligos, would be impossible due to the lack of a gold standard for biological data. Since such benchmarking would have been outside the scope of this study—especially for the many modifications considered here—we have refrained from doing so; instead we refer to a recently published systematic review for m⁶A base caller [70].

Furthermore, we call for an extensive and coordinated community effort to provide ground-truth data for establishing a gold-standard database for various tissue types, to better validate *in silico* predictions and, thus, to better train and compare the various tools for predicting modification detection.

Together with ModiDeC, we offer all our data as training and validation data for other tools and classifiers. We are providing an extensive database of carefully designed high-quality training and validation data for various RNA modifications, together with matching IVT data for synthetic data and for HEK293 cell lines. In detail, we provide training and test datasets for both SQK-RNA002 (Supplementary Figs S4 and S5) and SQK-RNA004 kits (Fig. 1 and Supplementary Fig. S3), and the datasets for the RNA oligos with variable complexity (Fig. 2, SQK-RNA004 kits) and HEK293 IVT reads (Fig. 4, SQK-RNA004 kits).

Although ModiDeC can be used in various scenarios, including human peripheral blood, our neural network model is limited in that only 40 motifs were generated for the initial model training. This limits the tool's utility in comprehensive genome-wide screens for routine diagnostic purposes despite its ability to extract features outside the training pool. To offer the possibility of personalized training for individual research questions, we developed a GUI (including an Epi2ME pipeline) that allows the user to retrain ModiDeC to search for user-defined specific sequences or motifs. We also see great potential in that our data pool will be a seed for further community-driven data generation. With the help of the GUI, any lab interested in a specific region in the transcriptome of a species can easily order matching oligos and train ModiDeC on this position or the corresponding motif, and then make it available to the community.

We argue that the ease of extension and customization of ModiDeC by anyone makes it a valuable tool for studies on RNA modification in biology and medicine. Also, if combined with models optimized for genome-wide screenings, ModiDeC could be a complementary high-precision validation tool.

Acknowledgements

[equal]), Johannes Miedema (Data curation [supporting], Formal analysis [supporting], Software [supporting], Visualization [supporting]), Stefan Pastore (Formal analysis [supporting], Methodology [supporting], Software [supporting]), Lioba Lehmann (Investigation [supporting], Visualization [supporting]), Anna Wierczeiko (Data curation [supporting], Investigation [supporting]), Johannes Friedrich (Data curation [supporting], Resources [supporting]), Lukas Walz (Data curation [supporting], Resources [supporting]), Marko Jörg (Data curation [supporting], Resources [supporting]), Tamer Butto (Data curation [supporting], Resources [supporting], Writing—review & editing [supporting]), Kristina Friedland (Investigation [supporting], Writing—original draft [supporting], Writing—review & editing [supporting]), Mark Helm (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Resources [equal], Supervision [equal], Writing—review & editing [supporting]), and Susanne Gerber (Conceptualization [equal], Funding acquisition [equal], Investigation [supporting], Methodology [supporting], Project administration [equal], Resources [equal], Supervision [equal], Writing—original draft [supporting], Writing—review & editing [lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Data availability

Training and test data sets for SQK-RNA002 kit, SQK-RNA004 kit, the data sets for the RNA oligos with variable complexity, and HEK293 IVT reads are available at the ENA database with the following ID projects: PRJEB88778. We also provide the full reference used to analyze the data (see supplementary_File_2_RNA_sequence). The neural network and the GUI codes for ModiDeC are available at Zenodo (DOI:10.5281/zenodo.15719181) and GitHub: https://github.com/mem3nto0/ModiDeC-RNA-modification-classifier, where tutorials for retraining and/or using ModiDeC are also provided. ModiDeC can also be found in the ONT platform Epi2ME with the names "ModiDeC data curation," "ModiDeC training," and "ModiDeC analysis."

References

1. Lucas MC, Novoa EM. Long-read sequencing in the era of epigenomics and epitranscriptomics. *Nat Methods* 2023;**20**:25–9.https://doi.org/10.1038/s41592-022-01724-8
2. Roundtree IA, Evans ME, Pan T *et al.* Dynamic RNA modifications in gene expression regulation. *Cell* 2017;**169**:1187–200. https://doi.org/10.1016/j.cell.2017.05.045
3. Boccaletto P, Machnicka MA, Purta E *et al.* MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res* 2018;**46**:D303–7. https://doi.org/10.1093/nar/gkx1030
4. Delaunay S, Helm M, Frye M. RNA modifications in physiology and disease: towards clinical applications. *Nat Rev Genet* 2024;**25**:104–22. https://doi.org/10.1038/s41576-023-00645-2
5. McMahon M, Forester C, Buffenstein R. Aging through an epitranscriptomic lens. *Nat Aging* 2021;**1**:335–46. https://doi.org/10.1038/s43587-021-00058-y
6. Shi H, Chai P, Jia R *et al.* Novel insight into the regulatory roles of diverse RNA modifications: re-defining the bridge between transcription and translation. *Mol Cancer* 2020;**19**:78. https://doi.org/10.1186/s12943-020-01194-6
7. Perez MF, Lehner B. Intergenerational and transgenerational epigenetic inheritance in animals. *Nat Cell Biol* 2019;**21**:143–51. https://doi.org/10.1038/s41556-018-0242-9
8. Jonkhout N, Tran J, Smith M *et al.* The RNA modification landscape in human disease. RNA 2017;**23**:1754–69. https://doi.org/10.1261/rna.063503.117
9. Wang X, Lu Z, Gomez A *et al.* N^6-Methyladenosine-dependent regulation of messenger RNA stability. *Nature* 2014;**505**:117–20. https://doi.org/10.1038/nature12730
10. Flamand MN, Meyer KD. m^{6}A and YTHDF proteins contribute to the localization of select neuronal mRNAs. *Nucleic Acids Res* 2022;**50**:4464–83. https://doi.org/10.1093/nar/gkac251
11. Jiang X, Liu B, Nie Z *et al.* The role of m^{6}A modification in the biological functions and diseases. *Signal Transduct Target Ther* 2021;**6**:74. https://doi.org/10.1038/s41392-020-00450-x
12. Helm M, Motorin Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat Rev Genet* 2017;**18**:275–91. https://doi.org/10.1038/nrg.2016.169
13. Carlile TM, Rojas-Duran MF, Zinshteyn B *et al.* Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* 2014;**515**:143–6. https://doi.org/10.1038/nature13802
14. Gabay O, Shoshan Y, Kopel E *et al.* Landscape of adenosine-to-inosine RNA recoding across human tissues. *Nat Commun* 2022;**13**:1184. https://doi.org/10.1038/s41467-022-28841-4
15. Srinivasan S, Torres AG, Ribas de Pouplana L. Inosine in biology and disease. *Genes* 2021;**12**:600. https://doi.org/10.3390/genes12040600
16. Feng Y-J, You X-J, Ding J-H *et al.* Identification of inosine and 2′-O-methylinosine modifications in yeast messenger RNA by liquid chromatography–tandem mass spectrometry analysis. *Anal Chem* 2022;**94**:4747–55. https://doi.org/10.1021/acs.analchem.1c05292
17. Safra M, Sas-Chen A, Nir R *et al.* The m^{1}A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* 2017;**551**:251–5. https://doi.org/10.1038/nature24456
18. Jörg M, Plehn JE, Kristen M *et al.* N^1-Methylation of adenosine (m^{1}A) in ND5 mRNA leads to complex I dysfunction in

Alzheimer's disease. *Mol Psychiatry* 2024;**29**:1427–39. https://doi.org/10.1038/s41380-024-02421-y

19. Hu L, Liu S, Peng Y *et al.* m^{6}A RNA modifications are measured at single-base resolution across the mammalian transcriptome. *Nat Biotechnol* 2022;**40**:1210–9. https://doi.org/10.1038/s41587-022-01243-z

20. Boulias K, Greer EL. Biological roles of adenine methylation in RNA. *Nat Rev Genet* 2023;**24**:143–60. https://doi.org/10.1038/s41576-022-00534-0

21. Hewel C, Wierczeiko A, Miedema J *et al.* Direct RNA sequencing (RNA004) enables improved transcriptome assessment and tracking of RNA modifications for medical applications. bioRxiv, https://www.biorxiv.org/content/10.1101/2024.07.25.605188v3, 2 January 2025, preprint: not peerreviewed.

22. Zhang Z, Luo K, Zou Z *et al.* Genetic analyses support the contribution of mRNA N^6-methyladenosine (m^{6}A) modification to human disease heritability. *Nat Genet* 2020;**52**:939–49. https://doi.org/10.1038/s41588-020-0644-z

23. Delaunay S, Frye M. RNA modifications regulating cell fate in cancer. *Nat Cell Biol* 2019;**21**:552–9. https://doi.org/10.1038/s41556-019-0319-0

24. Suzuki T. The expanding world of tRNA modifications and their disease relevance. *Nat Rev Mol Cell Biol* 2021;**22**:375–92. https://doi.org/10.1038/s41580-021-00342-0

25. Begik O, Lucas MC, Liu H *et al.* Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. *Genome Biol* 2020;**21**:97. https://doi.org/10.1186/s13059-020-02009-z

26. Frye M, Harada BT, Behm M *et al.* RNA modifications modulate gene expression during development. *Science* 2018;**361**:1346–9. https://doi.org/10.1126/science.aau1646

27. Meyer KD, Saletore Y, Zumbo P *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 2012;**149**:1635–46. https://doi.org/10.1016/j.cell.2012.05.003

28. Koh CWQ, Goh YT, Goh WSS. Atlas of quantitative single-base-resolution N^6-methyl-adenine methylomes. *Nat Commun* 2019;**10**:5636. https://doi.org/10.1038/s41467-019-13561-z

29. Carlile TM, Rojas-Duran MF, Gilbert WV. Pseudo-seq: genome-wide detection of pseudouridine modifications in RNA. *Methods Enzymol* 2015;**560**:219–45. https://doi.org/10.1016/bs.mie.2015.03.011

30. Linder B, Grozhik AV, Olarerin-George AO *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 2015;**12**:767–72. https://doi.org/10.1038/nmeth.3453

31. Liu C, Sun H, Yi Y *et al.* Absolute quantification of single-base m^{6}A methylation in the mammalian transcriptome using GLORI. *Nat Biotechnol* 2023;**41**:355–66. https://doi.org/10.1038/s41587-022-01487-9

32. Anreiter I, Mir Q, Simpson JT *et al.* New twists in detecting mRNA modification dynamics. *Trends Biotechnol* 2021;**39**:72–89. https://doi.org/10.1016/j.tibtech.2020.06.002

33. Oberdoeffer S, Gilbert WV. All the sites we cannot see: sources and mitigation of false negatives in RNA modification studies. *Nat Rev Mol Cell Biol* 2025;**26**:237–48. https://doi.org/10.1038/s41580-024-00784-2

34. Jain M, Abu-Shumays R, Olsen HE *et al.* Advances in Nanopore direct RNA sequencing. *Nat Methods* 2022;**19**:1160–4. https://doi.org/10.1038/s41592-022-01633-w

35. Wang Y, Zhao Y, Bollas A *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 2021;**39**:1348–65. https://doi.org/10.1038/s41587-021-01108-x

36. Amarasinghe SL, Su S, Dong X *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**:30. https://doi.org/10.1186/s13059-020-1935-5

37. Jenjaroenpun P, Wongsurawat T, Wadley TD *et al.* Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res* 2021;**49**:e7. https://doi.org/10.1093/nar/gkaa620

38. Stephenson W, Razaghi R, Busan S *et al.* Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell Genomics* 2022;**2**:100097. https://doi.org/10.1016/j.xgen.2022.100097

39. Begik O, Lucas MC, Pryszcz LP *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol* 2021;**39**:1278–91. https://doi.org/10.1038/s41587-021-00915-6

40. Nguyen TA, Heng JWJ, Kaewsapsak P *et al.* Direct identification of A-to-I editing sites with nanopore native RNA sequencing. *Nat Methods* 2022;**19**:833–44. https://doi.org/10.1038/s41592-022-01513-3

41. Spangenberg J, Mündnich S, Anne B *et al.* The RMaP challenge of predicting RNA modifications by nanopore sequencing. *Commun Chem* 2025;**8**:115. https://doi.org/10.1038/s42004-025-01507-0

42. Begik O, Mattick JS, Novoa EM. Exploring the epitranscriptome by native RNA sequencing. *RNA* 2022;**28**:1430–9. https://doi.org/10.1261/rna.079404.122

43. Tavakoli S, Nabizadeh M, Makhamreh A *et al.* Semi-quantitative detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing. *Nat Commun* 2023;**14**:334. https://doi.org/10.1038/s41467-023-35858-w

44. Furlan M, Delgado-Tejedor A, Mulroney L *et al.* Computational methods for RNA modification detection from Nanopore direct RNA sequencing data. *RNA Biol* 2021;**18**:31–40. https://doi.org/10.1080/15476286.2021.1978215

45. Pratanwanich PN, Yao F, Chen Y *et al.* Identification of differential RNA modifications from Nanopore direct RNA sequencing with xPore. *Nat Biotechnol* 2021;**39**:1394–402. https://doi.org/10.1038/s41587-021-00949-w

46. Abebe JS, Price AM, Hayer KE *et al.* DRUMMER—rapid detection of RNA modifications through comparative nanopore sequencing. *Bioinformatics* 2022;**38**:3113–5. https://doi.org/10.1093/bioinformatics/btac274

47. Hendra C, Pratanwanich PN, Wan YK *et al.* Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat Methods* 2022;**19**:1590–8. https://doi.org/10.1038/s41592-022-01666-1

48. Qin H, Ou L, Gao J *et al.* DENA: training an authentic neural network model using nanopore sequencing data of Arabidopsis transcripts for detection and quantification of N^6-methyladenosine on RNA. *Genome Biol* 2022;**23**:25. https://doi.org/10.1186/s13059-021-02598-3

49. Chan A, Naarmann-de Vries IS, Scheitl CPM *et al.* Detecting m^{6}A at single-molecular resolution via direct RNA sequencing and realistic training data. *Nat Commun* 2024;**15**:3323. https://doi.org/10.1038/s41467-024-47661-2

50. Hassan D, Acevedo D, Daulatabad SV *et al.* Penguin: a tool for predicting pseudouridine sites in direct RNA nanopore sequencing data. *Methods* 2022;**203**:478–87. https://doi.org/10.1016/j.ymeth.2022.02.005

51. Burdick JT, Comai A, Bruzel A *et al.* Nanopore-based direct sequencing of RNA transcripts with 10 different modified nucleotides reveals gaps in existing technology. *G3* 2023;**13**:jkad200. https://doi.org/10.1093/g3journal/jkad200

52. Acera Mateos P, Sethi A, Ravindran A *et al.* Prediction of m6A and m5C at single-molecule resolution reveals a transcriptome-wide co-occurrence of RNA modifications. *Nat Commun* 2024;**15**:3899. https://doi.org/10.1038/s41467-024-47953-7

53. Wu Y, Shao W, Yan M *et al.* Transfer learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing. *Nat Commun* 2024;**15**:4049. https://doi.org/10.1038/s41467-024-48437-4

54. Lucas MC, Pryszcz LP, Medina R *et al.* Quantitative analysis of tRNA abundance and modifications by nanopore RNA

sequencing. *Nat Biotechnol* 2024;**42**:72–86. https://doi.org/10.1038/s41587-023-01743-6

55. Milenkovic I, Cruciani S, Llovera L *et al.* Epitranscriptomic rRNA fingerprinting reveals tissue-of-origin and tumor-specific signatures. *Mol Cell* 2025;**85**:177–90. https://doi.org/10.1101/2024.10.03.616461

56. Begik O, Diensthuber G, Liu H *et al.* Nano3P-Seq: transcriptome-wide analysis of gene expression and tail dynamics using end-capture nanopore cDNA sequencing. *Nat Methods* 2023;**20**:75–85. https://doi.org/10.1038/s41592-022-01714-w

57. Huang S, Wylder AC, Pan T. Simultaneous nanopore profiling of mRNA m^{6}A and pseudouridine reveals translation coordination. *Nat Biotechnol* 2024;**42**:1831–5. https://doi.org/10.1038/s41587-024-02135-0

58. Schartel L, Jann C, Wierczeiko A *et al.* Selective RNA pseudouridinylation *in situ* by circular gRNAs in designer organelles. *Nat Commun* 2024;**15**:9177. https://doi.org/10.21203/rs.3.rs-4756705/v1

59. Pastore S, Wacheul L, Lehmann L *et al.* Mapping human pre rRNA processing and modification at single nucleotide resolution. bioRxiv, https://doi.org/10.1101/2025.03.01.640970, 28 May 2025, preprint: not peer reviewed.

60. Song Y. mRNA methylation de-mystified. *Nat Chem Biol* 2023;**19**:252. https://doi.org/10.1038/s41589-023-01290-w

61. Samarakoon H, Wan YK, Parameswaran S *et al.* Leveraging basecaller's move table to generate a lightweight *k*-mer model. *Bioinformatics* 2024;**41**:btaf111. https://doi.org/10.1101/2024.06.30.601452

62. Taoka M, Nobe Y, Yamaki Y *et al.* Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res* 2018;**46**:9289–98. https://doi.org/10.1093/nar/gky811

63. Shoemaker R, Huang M-F, Wu Y-S *et al.* Decoding the molecular symphony: interactions between the m^{6}A and p53 signaling pathways in cancer. *NAR Cancer* 2024;**6**:zcae037. https://doi.org/10.1093/narcan/zcae037

64. Hernández Borrero LJ, El-Deiry WS. Tumor suppressor p53: biology, signaling pathways, and therapeutic targeting. *Biochim Biophys Acta Rev Cancer* 2021;**1876**:188556. https://doi.org/10.1016/j.bbcan.2021.188556

65. Wang H, Guo M, Wei H *et al.* Targeting p53 pathways: mechanisms, structures and advances in therapy. *Signal Transduct Target Ther* 2023;**8**:92. https://doi.org/10.1038/s41392-023-01347-1

66. Xuan J, Chen L, Chen Z *et al.* RMBase v3.0: decode the landscape, mechanisms and functions of RNA modifications. *Nucleic Acids Res* 2024;**52**:D273–84. https://doi.org/10.1093/nar/gkad1070

67. Liang Z, Ye H, Ma J *et al.* m^{6}A-Atlas v2.0: updated resources for unraveling the N^6-methyladenosine (m^{6}A) epitranscriptome among multiple species. *Nucleic Acids Res* 2024;**52**:D194–202. https://doi.org/10.1093/nar/gkad691

68. Li X, Xiong X, Zhang M *et al.* Base-resolution mapping reveals distinct m^{1}A methylome in nuclear- and mitochondrial-encoded transcripts. *Mol Cell* 2017;**68**:993–1005. https://doi.org/10.1016/j.molcel.2017.10.019

69. Lin T-Y, Kleemann L, Jeżowski J *et al.* The molecular basis of tRNA selectivity by human pseudouridine synthase 3. *Mol Cell* 2024;**84**:2472–89. https://doi.org/10.1016/j.molcel.2024.06.013

70. Maestri S, Furlan M, Mulroney L *et al.* Benchmarking of computational methods for m6A profiling with nanopore direct RNA sequencing. *Brief Bioinform* 2024;**25**:bbae001. https://doi.org/10.1093/bib/bbae001