Validation Strategies for the Detection of m^5C by Nanopore Direct RNA Sequencing

Dissertation

zur Erlangung des Grades

»Doktor der Naturwissenschaften«

im Promotionsfach Pharmazie

am Fachbereich Chemie, Pharmazie, Geographie und Geowissenschaften

der Johannes Gutenberg-Universität Mainz

Zeynep Özrendeci

geboren in Fethiye

Mainz, Februar 2026

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

Dekanin: [redacted]

1. Berichterstatter: [redacted]

2. Berichterstatterin: [redacted]

Datum der mündlichen Prüfung: 28. April 2026

D77 (Dissertation Mainz)

Die vorliegende Arbeit wurde am Institut für Pharmazeutische und Biomedizinische Wissenschaften der Johannes Gutenberg-Universität in Mainz zur Erlangung des Grades »Doktor der Naturwissenschaften« unter der Betreuung von [redacted] angefertigt.

1. Berichterstatter

[redacted]

Fachbereich Chemie, Pharmazie, Geographie und Geowissenschaften

Institut für Pharmazeutische und Biomedizinische Wissenschaften

Johannes Gutenberg-Universität Mainz

2. Berichterstatterin

[redacted]

Fachbereich Chemie, Pharmazie, Geographie und Geowissenschaften

Institut für Pharmazeutische und Biomedizinische Wissenschaften

Johannes Gutenberg-Universität Mainz

Hiermit versichere ich eidesstattlich:

1. Ich habe die jetzt als Dissertation vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel (Literatur, Apparaturen, Material) in der Arbeit angegeben.

2. Ich habe oder hatte die jetzt als Dissertation vorgelegte Arbeit nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

3. Ich hatte weder die jetzt als Dissertation vorgelegte Arbeit noch Teile davon bei einer anderen Fakultät bzw. einem anderen Fachbereich als Dissertation eingereicht.

Ort, Datum

Unterschrift

Acknowledgements

Contents

Acknowledgements ..... **iv**

ABSTRACT ..... **viii**

ZUSAMMENFASSUNG ..... **ix**

List of Abbreviations ..... **x**

1. Introduction **1**

1.1 The Central Dogma of Molecular Biology and the Emergence of RNA as a Regulatory Hub *1*

1.1.1 From Unidirectional Flow to Regulatory Complexity 1

1.1.2 RNA Structure and the Significance of Single-Nucleotide Chemistry 2

1.2 RNA Modifications: 5-Methylcytidine and 2'-O-Methylation *5*

1.2.1 A Historical Overview of RNA modifications 5

1.2.2 5-Methylcytidine Biochemistry: Structural and Functional Roles 5

1.2.3 The Enzymatic Machinery of m^5C 6

1.2.4 Stoichiometry and Biological Dynamics of m^5C 9

1.2.5 2’-O-Methylation: Chemistry, Biology, and Relevance to Viral RNA 10

1.3 Conventional RNA Modification Detection Methods *11*

1.3.1 LC-MS/MS (Liquid Chromatography-Mass Spectrometry) 11

1.3.2 Bisulfite Sequencing (BS-seq) 11

1.3.3 Antibody-Based and Enzyme-Assisted Methods 12

1.3.4 Detection Methods for 2'-O-Methylation 13

1.3.5 Fundamental Limitations of Conventional Methods 15

1.4 Nanopore Direct RNA Sequencing (DRS) *15*

1.4.1 Physical Principles and Fundamental Terminology 16

1.4.2 Modification Detection via Nanopore DRS 18

1.4.3 RNA Modification Detection Tools Using Nanopore DRS 19

1.4.4 Limitations of Classical Nanopore Approaches and the Transition to Deep Learning 21

1.5 Deep Learning in Nanopore Direct RNA Sequencing *22*

1.5.1 Fundamentals of Deep Learning 22

1.5.2 Neural Network Architectures for Nanopore Signal Analysis 24

1.5.3 Deep Learning-Powered Tools for RNA Modification Detection 26

1.5.4 The Challenge of m^5C Detection by Nanopore DRS 27

1.5.5 Training Strategies and Ground Truth Generation 28

1.5.6 Hard Negative Mining: A Principled Framework for Decision Boundary Refinement 30

1.5.7 The Validation Gap: From Discovery to Confirmation 32

2. Motivation and Objectives **34**

3. Results and Discussion **36**

3.1 Ligation Strategies for Site-Specifically Modified RNA *36*

3.1.1 Validation of SARS-CoV-2 2'-O-Methylation Candidates by RiboMethSeq 36

3.1.2 Chemical Capping and Application to Translational Studies 48

3.2 ModiCal: Calibration-Based m^5C Detection by Nanopore DRS *54*

3.2.1 Rationale for Model System Selection 54

3.2.3 Iterative Calibration: From Widespread Overcalling to Site-Specific Precision 59

3.2.4 Dual-Site Calibration: Extending Detection to Multiple m^5C Sites on Yeast 25S rRNA 65

3.2.5 Cross-RNA Transferability: Application of the ModiCal Workflow to Dengue Virus Genomic RNA 71

3.2.6 A Generalized Decision Framework for ModiCal Implementation 74

4. Conclusion and Perspectives **76**

4.1 Summary of Ligation Strategies and Biochemical Foundations *76*

4.2 ModiCal: Calibration-Based Modification Detection *76*

4.3 Limitations and Future Perspectives *77*

4.4 Concluding Remarks *78*

5. Materials and Methods **79**

5.1 Materials *79*

5.1.1 Instruments 79

5.1.2 Chemicals and Reagents 79

5.1.3 Enzymes 80

5.1.4 Kits 81

5.1.5 Oligonucleotides 81

5.1.6 Plasmid Vectors 82

5.1.7 Yeast Strains 83

5.1.8 Buffers and Solutions 83

5.2 Methods *83*

5.2.1 RNA Preparation 83

5.2.2 *In Vitro* Translation 85

5.2.3 Fluorescence Detection 86

5.2.4 RiboMethSeq 86

5.2.5 Bisulfite Sequencing 86

5.2.6 Nanopore Direct RNA Sequencing (DRS) 87

5.2.7 ModiCal Calibration Workflow 87

5.2.8 Bioinformatic Analysis 88
5.2.9 Dorado Modification Calling 88
5.2.10 Computational Resources 89
5.2.11 Data and Code Availability 89
6. Bibliography 90
List of Publications and Presentations 109
Appendix 110
A.1 Oligonucleotide Sequences 110
A.1.1 Long Ligation Constructs 110
A.1.2 Short Ligation Constructs 111
A.2 Plasmid Insert Sequences 111
A.3 Software and Computational Resources 114
A.3.1 Software Packages 114
A.3.2 Computational Resources 114
A.3.3 Use of Artificial Intelligence and Writing Tools 114
Curriculum Vitae 115

ABSTRACT

RNA modifications play essential roles in gene expression, yet their reliable detection at single-site resolution remains a significant analytical challenge. Among these modifications, 5-methylcytidine (m^5C) is particularly difficult to detect due to its subtle ionic current perturbation and typically low stoichiometry. This challenge persists even with nanopore direct RNA sequencing (DRS), the current cutting-edge platform for native RNA analysis. Despite advances in deep learning-based approaches, existing methods generate false-positive rates that preclude reliable site-specific interpretation. This thesis describes ModiCal (ModiDeC-based Calibration), a calibration-driven framework that repurposes a neural network classifier into a high-precision m^5C validation tool. The work proceeded through two phases. In the first, different splint-mediated ligation strategies for generating site-specific modified RNAs were explored, including a chemical capping protocol for potential application in modification detection. Short ligation constructs were confirmed to be fully compatible with nanopore DRS. Attempted validation of putative 2′-O-methylation sites on SARS-CoV-2 transcripts by RiboMethSeq revealed systematic false-positive signals that were indistinguishable from genuine modification, exposing a fundamental limitation of cleavage-based detection and driving the transition to nanopore-based approaches. In the second phase, the ModiCal workflow was established using yeast 25S ribosomal RNA. A three-step calibration strategy was implemented: baseline training on synthetic ground-truth RNA, bulk false-positive suppression using *in vitro* transcribed reference RNA, and iterative single-site refinement. This approach reframes false-positive predictions as diagnostic readouts of misaligned decision boundaries. The calibrated model achieved background-free, enzyme-dependent detection of both characterized m^5C sites on yeast rRNA. Applied to dengue virus genomic RNA, it confirmed a low-stoichiometry m^5C site with zero false positives across the entire viral genome. Quantitative accuracy was further validated by stoichiometry titration. ModiCal establishes a map-and-validate paradigm in which discovery tools nominate candidates and calibration-based validation provides site-specific confirmation. The principles demonstrated here are modification-agnostic and applicable beyond m^5C.

ZUSAMMENFASSUNG

RNA-Modifikationen spielen eine zentrale Rolle in der Genexpression, doch ihr zuverlässiger Nachweis auf Einzelpositionsebene stellt nach wie vor eine erhebliche analytische Herausforderung dar. Unter diesen Modifikationen ist 5-Methylcytidin (m^5C) aufgrund seiner subtilen Ionenstrom-Perturbation und typischerweise niedrigen Stöchiometrie besonders schwer zu detektieren. Diese Herausforderung besteht selbst bei Nanopore-Direct-RNA-Sequencing (DRS), der derzeit fortschrittlichsten Plattform für die Analyse nativer RNA. Trotz Fortschritten bei Deep-Learning-basierten Ansätzen erzeugen bestehende Methoden Falsch-Positiv-Raten, die eine zuverlässige positionsspezifische Interpretation ausschließen. Diese Arbeit beschreibt ModiCal (ModiDeC-basierte Kalibrierung), ein kalibrierungsgetriebenes Framework, das einen neuronalen Klassifikator in ein hochpräzises m^5C-Validierungswerkzeug überführt. Die Arbeit gliedert sich in zwei Phasen. In der ersten wurden verschiedene Splint-vermittelte Ligationsstrategien zur Herstellung positionsspezifisch modifizierter RNAs erprobt, einschließlich eines chemischen Capping-Protokolls zur potenziellen Anwendung in der Modifikationsdetektion. Kurze Ligationskonstrukte erwiesen sich als vollständig kompatibel mit Nanopore-DRS. Die versuchte Validierung potenzieller 2′-O-Methylierungsstellen auf SARS-CoV-2-Transkripten mittels RiboMethSeq ergab systematische Falsch-Positiv-Signale, die von echten Modifikationen nicht unterscheidbar waren. Dies legte eine grundlegende Limitation spaltungsbasierter Detektion offen und veranlasste den Wechsel zu Nanopore-basierten Ansätzen. In der zweiten Phase wurde der ModiCal-Workflow an ribosomaler Hefe-25S-RNA etabliert. Eine dreistufige Kalibrierungsstrategie wurde implementiert: Baseline-Training auf synthetischer Ground-Truth-RNA, Bulk-Unterdrückung von Falsch-Positiven mittels in-vitro-transkribierter Referenz-RNA und iterative Einzelsite-Verfeinerung. Dieser Ansatz interpretiert Falsch-Positiv-Vorhersagen als diagnostische Hinweise auf fehlkalibrierte Entscheidungsgrenzen. Das kalibrierte Modell erzielte hintergrundfreie, enzymabhängige Detektion beider charakterisierten m^5C-Stellen der Hefe-rRNA. Angewandt auf Dengue-Virus-Genom-RNA bestätigte es eine m^5C-Stelle niedriger Stöchiometrie bei null Falsch-Positiven über das gesamte virale Genom. Die quantitative Genauigkeit wurde durch Stöchiometrie-Titration validiert. ModiCal etabliert ein Map-and-Validate-Paradigma, bei dem Discovery-Tools Kandidaten nominieren und kalibrierungsbasierte Validierung positionsspezifische Bestätigung liefert. Die hier gezeigten Prinzipien sind modifikationsagnostisch und über m^5C hinaus anwendbar.

List of Abbreviations

Abbreviation	Definition
ac^4C	N4-acetylcytidine
ALYREF	ALY/REF export factor
Am	2'-O-methyladenosine
APS	Ammonium persulfate
ARCA	Anti-reverse cap analog
ATP	Adenosine triphosphate
AU-ROC	Area under the receiver operating characteristic curve
AUC	Area under the curve
AUPRC	Area under the precision-recall curve
BAM	Binary Alignment Map
cDNA	Complementary DNA
CHEUI	CHEmical modIfication detection on nanopore sequencing
Cm	2'-O-methylcytidine
CNN	Convolutional neural network
CPU	Central processing unit
Cy5	Cyanine 5
DENV	Dengue virus
DL	Deep learning
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNMT2	DNA methyltransferase 2
dNTP	Deoxynucleotide triphosphate
DRS	Direct RNA sequencing
dsDNA	Double-stranded DNA
DTT	Dithiothreitol
DTW	Dynamic time warping
eGFP	Enhanced green fluorescent protein
ELIGOS	Epitranscriptional Landscape Inferring from Glitches of Oxford Nanopore Signals
ENA	European Nucleotide Archive
F1	Fragment 1 (IVT fragment)
F2	Fragment 2 (synthetic oligoribonucleotide)
F3	Fragment 3 (IVT fragment)
f^5C	5-formylcytidine
FDR	False discovery rate

FL	Full-length
FN	False negative
FP	False positive
Gm	2'-O-methylguanosine
GMP	Guanosine monophosphate
GPU	Graphics processing unit
gRNA	Genomic RNA
GTP	Guanosine triphosphate
hm^5C	5-hydroxymethylcytidine
HMM	Hidden Markov model
HPLC	High-performance liquid chromatography
IFIT	Interferon-induced protein with tetratricopeptide repeats
IGV	Integrative Genomics Viewer
IP	Immunoprecipitation
IRES	Internal ribosome entry site
IUPAC	International Union of Pure and Applied Chemistry
IVT	*In vitro* transcription
LC-MS/MS	Liquid chromatography–tandem mass spectrometry
lncRNA	Long non-coding RNA
LSTM	Long short-term memory
m^1A	N1-methyladenosine
m^5C	5-methylcytidine
m^6A	N6-methyladenosine
m^7G	7-methylguanosine
miCLIP	Methylation individual-nucleotide-resolution crosslinking and immunoprecipitation
MIL	Multiple instance learning
MINES	m^6A Identification using Nanopore sequencing
ML	Machine learning
ModiCal	ModiDeC-based Calibration
ModiDeC	Modification Detection Classifier
mRNA	Messenger RNA
MTERF4	Mitochondrial transcription termination factor 4
NGS	Next-generation sequencing
Nm	2'-O-methylation
NOP2	Nucleolar protein 2
NSUN	NOL1/NOP2/SUN domain protein family
nt	Nucleotide(s)
NTP	Nucleoside triphosphate
OHEM	Online hard example mining

ONT	Oxford Nanopore Technologies
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
pDNA	Plasmid DNA
PNK	Polynucleotide kinase
RCM1	rRNA cytosine methyltransferase 1 (yeast NSUN5 ortholog)
RIP	RNA immunoprecipitation
RNA	Ribonucleic acid
RNA-BS-seq	RNA bisulfite sequencing
RNN	Recurrent neural network
ROC	Receiver operating characteristic
RP-HPLC	Reversed-phase high-performance liquid chromatography
RRL	Rabbit reticulocyte lysate
rRNA	Ribosomal RNA
RTGE	Real-time gel elution
RTL-P	Reverse transcription at low dNTP concentrations followed by PCR
S1-NTD	S1 N-terminal domain (SARS-CoV-2 Spike protein)
SAH	S-adenosyl-L-homocysteine
SAM	S-adenosyl-L-methionine
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SF3B1	Splicing factor 3B subunit 1
T4	Bacteriophage T4
T7	Bacteriophage T7
TAC-seq	TET-assisted chemical labeling sequencing
TBE	Tris-borate-EDTA
TEMED	Tetramethylethylenediamine
TET	Ten-eleven translocation (dioxygenase family)
TN	True negative
TP	True positive
tRNA	Transfer RNA
Um	2'-O-methyluridine
UTR	Untranslated region
UV	Ultraviolet
VRAM	Video random-access memory
WT	Wild type
YBX1	Y-box binding protein 1

1. Introduction

1.1 The Central Dogma of Molecular Biology and the Emergence of RNA as a Regulatory Hub

1.1.1 From Unidirectional Flow to Regulatory Complexity

The conceptual framework of molecular biology was shaped in 1958 when Francis Crick articulated the "Central Dogma", proposing a rigorous, unidirectional flow of genetic information. Deoxyribonucleic acid (DNA) makes ribonucleic acid (RNA), and RNA makes protein (Crick, 1958). In this foundational model, RNA was viewed simply as a temporary carrier, responsible only for relaying genetic instructions from DNA to the protein synthesis machinery.

Subsequent discoveries, including reverse transcription (Baltimore, 1970; Temin & Mizutani, 1970), catalytic RNAs (Kruger et al., 1982; Guerrier-Takada et al., 1983), and the RNA World hypothesis (Gilbert, 1986), progressively dismantled this minimalist view, revealing RNA as an active participant in cellular regulation rather than a passive information shuttle.

The most profound expansion of the Central Dogma, however, lies in the realization that RNA is not merely a sequence-based messenger but a highly sophisticated, chemically modified molecule. The "passive mediator" role has been superseded by the concept of the epitranscriptome, a layer of post-transcriptional regulation mediated by over 170 distinct chemical modifications (Boccaletto et al., 2022; Sordyl et al., 2026). These modifications, such as the widely studied N6-methyladenosine (m^6A) and the focus of this thesis, 5-methylcytidine (m^5C), introduce a dynamic dimension to RNA function that is not encoded in the DNA template. Modern molecular biology now recognizes that RNA molecules undergo complex life cycles, spanning from alternative splicing and nuclear export to targeted degradation and translational control, all of which are finely tuned by their chemical state (Figure 1.1) (Roundtree et al., 2017; Frye et al., 2018).

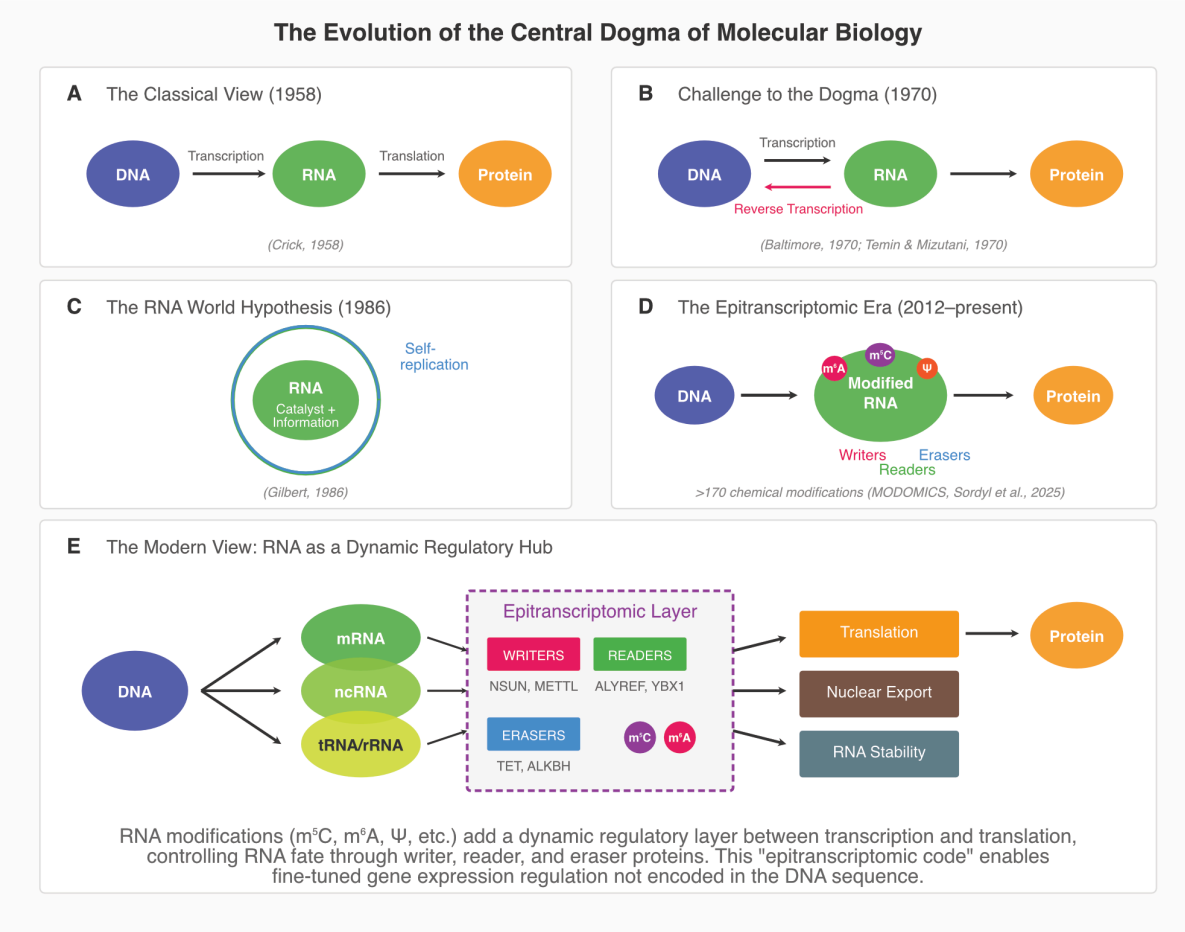


Figure 1.1 The Evolution of the Central Dogma of Molecular Biology. (A) The classical unidirectional model of genetic information flow. (B) The discovery of reverse transcriptase demonstrated that genetic information could flow from RNA back to DNA. (C) The RNA World hypothesis proposed that early life forms relied solely on RNA molecules. (D) The epitranscriptomic era revealed that RNA carries over 170 chemically distinct post-transcriptional modifications. (E) Contemporary model depicting RNA as a central regulatory hub.

1.1.2 RNA Structure and the Significance of Single-Nucleotide Chemistry

To appreciate how modifications alter RNA function, one must first consider the composition of the ribonucleotide, the fundamental building block of the transcriptome. Each nucleotide consists of three components. A phosphate group, a five-carbon ribose sugar, and a nitrogenous nucleobase (Figure 1.2). The four canonical nucleobases are divided into two chemical classes. The purines, adenine (A) and guanine (G), which contain a fused double-ring system, and the pyrimidines, cytosine (C) and uracil (U), which contain a single six-membered ring. Through Watson-Crick base pairing, adenine pairs with uracil and guanine pairs with cytosine, forming the hydrogen-bonded interactions that stabilize RNA secondary structures (Leontis & Westhof, 2001). While the phosphodiester backbone provides structural continuity, a critical distinction from DNA is the presence of a 2'-hydroxyl group on the ribose, which grants RNA its

characteristic chemical reactivity and susceptibility to hydrolysis (Helm & Motorin, 2017). It is the nitrogenous bases, however, that carry the primary genetic information. This four-letter alphabet is merely the primary canvas; the true functional diversity of RNA arises from the chemical diversification of these building blocks through over 170 distinct post-transcriptional modifications (Boccaletto et al., 2022; Sordyl et al., 2026).

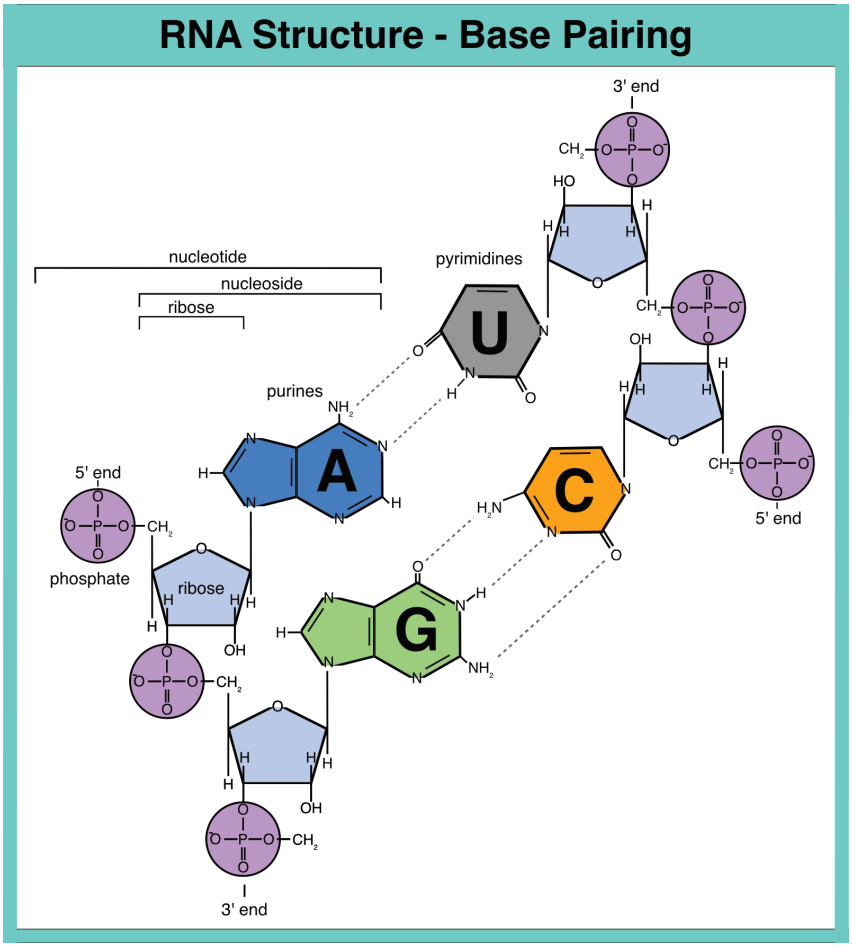


Figure 1.2. Structure of ribonucleotides and Watson-Crick base pairing in RNA. Each nucleotide consists of a phosphate group (purple), a ribose sugar (blue), and a nitrogenous nucleobase. The purine bases adenine (A, blue) and guanine (G, green) contain a fused double-ring system, while the pyrimidine bases uracil (U, gray) and cytosine (C, orange) contain a single six-membered ring. Dashed lines indicate canonical Watson-Crick hydrogen bonding between A-U and G-C base pairs. The hierarchical composition of a nucleotide, nucleoside, and ribose is indicated by brackets (left). The 5′ and 3′ directionality of each strand and the 2′-hydroxyl group (OH) on the ribose are labeled.

The functional versatility of RNA is intrinsically linked to its structural plasticity. Unlike the relatively uniform DNA double helix, RNA's single-stranded nature allows it to fold into complex secondary and tertiary architectures (Leontis & Westhof, 2001). Messenger RNA (mRNA) possesses intricate structural elements such as internal ribosome entry sites (IRES) and riboswitches that modulate translational efficiency (Hentze et al., 2018), while tRNA and rRNA represent the pinnacle of structural conservation. The L-shaped folding of tRNA is essential for aminoacyl-tRNA synthetase recognition, and the massive structure of rRNA forms the catalytic

core of the ribosome (Moore & Steitz, 2002). These highly structured RNAs are also the primary hosts for dense chemical modifications critical for maintaining their folding under physiological stress.

The chemical modification of a single atom can profoundly alter the physicochemical properties of the entire RNA molecule. For instance, the methylation of cytidine at the C5 position increases the hydrophobicity of the pyrimidine ring and enhances base-stacking interactions (Motorin & Helm, 2011). Similarly, 2'-O-methylation of the ribose sugar blocks the 2'-hydroxyl's participation in strand cleavage and alters local backbone flexibility, with consequences for both RNA stability and innate immune recognition (Daffis et al., 2010). These modifications exert their effects through three primary mechanisms: (i) structural tuning of thermodynamic stability and folding, as exemplified by m^5C stabilizing the tertiary fold of tRNA (Grosjean, 2009); (ii) serving as steric roadblocks or docking signals for RNA-binding proteins (Yang et al., 2017; Roundtree et al., 2017); and (iii) altering base-pairing dynamics, as in adenosine-to-inosine RNA editing (Nishikura, 2016).

Critically, the biological impact of a modification depends not only on its chemical identity but also on its exact position. In tRNA, m^5C at position 34 (the wobble position) directly influences translational fidelity, whereas m^5C at position 48 stabilizes the variable loop (Sordyl et al., 2026). Furthermore, the epitranscriptomic code exhibits stoichiometric heterogeneity. At any given site, only a fraction of transcripts may carry the modification. Within a single cellular environment, a specific cytidine residue on a given mRNA species is rarely methylated in 100% of the copies. Instead, it exists in a fractional state, and this fractional methylation is not biological noise; it represents a sophisticated layer of dosage control where sub-populations often follow divergent fates. The methylated fraction might be recognized by a reader protein, triggering its export to the cytoplasm, while the unmethylated fraction might remain in the nucleus or be targeted for degradation (Squires et al., 2012; Yang et al., 2017; Acera Mateos et al., 2024).

This chapter has established that RNA is far more than a passive messenger. The following sections will focus specifically on 5-methylcytidine and 2'-O-methylation, tracing their biochemistry and the molecular consequences of their presence in different RNA classes, before addressing the technological challenges of their detection.

1.2 RNA Modifications: 5-Methylcytidine and 2'-O-Methylation

1.2.1 A Historical Overview of RNA modifications

The history of RNA modifications is a journey from the discovery of atypical nucleotides in highly abundant RNAs to the identification of a pervasive regulatory code across the entire transcriptome. The field began in 1951 with the discovery of pseudouridine by Cohn and Volkin, often referred to as the "fifth nucleotide" due to its ubiquity (Cohn & Volkin, 1951). During the 1960s and 1970s, RNA modifications including m^5C in tRNA were largely viewed as static marks essential for structural integrity of housekeeping RNAs (Grosjean, 2009). The field entered a period of relative dormancy regarding mRNA, as it was widely assumed that the transient nature of messenger transcripts precluded the need for such complex chemical regulation.

The paradigm shifted substantially in the early 2010s, catalyzed by the convergence of high-throughput sequencing and antibody-based enrichment techniques. The landmark mapping of m^6A in mRNA revealed that modifications were non-randomly distributed, conserved, and dynamically regulated (Dominissini et al., 2012; Meyer et al., 2012), giving birth to the term "Epitranscriptomics" (Saletore et al., 2012). Today, the MODOMICS database lists over 170 distinct RNA modifications (Sordyl et al., 2026), recognized as a regulatory language governing every aspect of RNA biology (Roundtree et al., 2017; Delaunay et al., 2024). Among these, m^5C emerges as a particularly significant mark. While its presence in tRNA and rRNA has been documented for over half a century, its recent identification in mRNA and long non-coding RNAs has opened new frontiers in understanding cytidine methylation as a dynamic regulatory signal (Yang et al., 2017; Squires et al., 2012).

1.2.2 5-Methylcytidine Biochemistry: Structural and Functional Roles

The chemical identity of 5-methylcytidine is defined by the covalent addition of a methyl group to the fifth carbon of the cytosine ring (Figure 1.3), inducing significant changes in base-stacking energies and the pKa of the nitrogenous base (Motorin et al., 2010). In tRNAs, m^5C at positions 34, 40, 48, 49, and 50 stabilizes the tertiary fold and protects against endonucleolytic cleavage by stress-induced ribonucleases such as angiogenin (Blanco et al., 2014; Hussain et al., 2013; Sordyl et al., 2026). In rRNA, m^5C at position 2278 of the yeast 25S rRNA is located near the

peptidyl transferase center, where it fine-tunes the positioning of the aminoacyl-tRNA during translation elongation (Schosserer et al., 2015; Sharma et al., 2013).

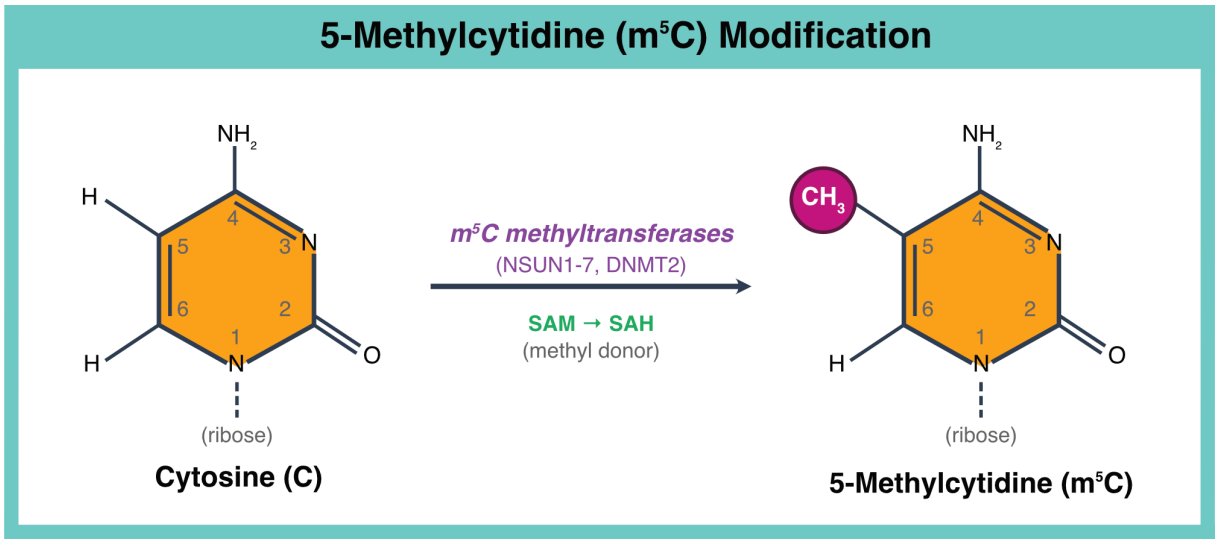


Figure 1.3. The 5-methylcytidine (m^5C) modification. Cytosine (C, left) is converted to 5-methylcytidine (m^5C, right) by the covalent addition of a methyl group (CH_3, magenta) to the C5 position of the pyrimidine ring. The reaction is catalyzed by m^5C methyltransferases of the NSUN family (NSUN1–7) and DNMT2, using S-adenosyl-L-methionine (SAM) as the methyl donor, which is converted to S-adenosyl-L-homocysteine (SAH) upon methyl transfer. Ring atom positions are numbered according to IUPAC convention. The dashed line indicates the N-glycosidic bond to the ribose sugar.

In mRNA and lncRNA, m^5C serves as a dynamic regulatory signal rather than a static structural element. It recruits the export adaptor ALYREF for nucleocytoplasmic transport, directly facilitating the efficient translocation of transcripts from the nucleus to the cytoplasm (Yang et al., 2017). Beyond export, m^5C influences mRNA stability by recruiting the reader protein YBX1, a process particularly vital during early embryonic development and cellular differentiation (Chen et al., 2019). Furthermore, during oxidative stress, the m^5C landscape on mRNA shifts, altering the stability of transcripts involved in metabolic homeostasis and pro-survival pathways (Huang et al., 2019; Delaunay & Frye, 2019). The biological necessity of precise m^5C placement is underscored by severe phenotypes. In the brain, the absence of specific m^5C marks leads to impaired neural stem cell differentiation and microcephaly (Khan et al., 2012), while in cancer, m^5C hypermethylation often occurs in oncogenic transcripts, promoting tumor progression and metastasis (Mei et al., 2020; Chen et al., 2021; Delazer et al., 2025).

1.2.3 The Enzymatic Machinery of m^5C

1.2.3.1 The Catalytic Mechanism

The enzymatic conversion of cytosine to m^5C is a multi-step process requiring a covalent enzyme-RNA intermediate, utilizing S-adenosyl-L-methionine (SAM) as the methyl donor (Figure 1.3) (Bohnsack et al., 2019; Höbartner et al., 2024). The reaction is initiated when a

conserved catalytic cysteine in the enzyme's active site performs a nucleophilic attack on the C6 position of the cytosine base, disrupting the C5-C6 double bond and enhancing the nucleophilicity of C5. The activated C5 carbon then attacks the labile methyl group of SAM, completing the transfer and producing S-adenosyl-L-homocysteine. A second conserved residue facilitates beta-elimination by deprotonating C5, restoring the C5-C6 double bond and releasing the modified m^5C base with regeneration of the active enzyme (Höbartner et al., 2024).

1.2.3.2 The Writers

The primary m^5C writers in eukaryotes are the NSUN (NOL1/NOP2/SUN domain) family proteins and DNMT2 (Table 1.1). Of particular relevance to this thesis are NSUN1 and NSUN5, which maintain the structural integrity of the large ribosomal subunit. In *Saccharomyces cerevisiae*, Nop2 (the NSUN1 ortholog) catalyzes the methylation of C2870 in the 25S rRNA, a site located near the peptidyl transferase center, and its absence causes significant defects in pre-rRNA processing and 60S subunit biogenesis (Sharma et al., 2013). Conversely, Rcm1 (the NSUN5 ortholog) modifies C2278 in the 25S rRNA; notably, the loss of Rcm1-mediated m^5C at this site reprograms the ribosome toward stress-response transcripts and extends yeast lifespan, illustrating that single-nucleotide chemistry can dictate organismal longevity (Schosserer et al., 2015). NSUN2 is the most prolific m^5C writer, modifying nearly all tRNAs at positions C34, C48, and C50, as well as targeting mRNA and vault RNA. Its dysregulation is central to Dubowitz-like Syndrome, where the loss of m^5C leads to tRNA fragmentation and neurodevelopmental failure, and its overexpression in squamous cell carcinoma drives tumor progression (Yang et al., 2017; Delazer et al., 2025). NSUN6 occupies a unique niche due to its stringent requirement for both sequence and structural context. Structural studies have revealed that NSUN6 acts as a molecular ruler, measuring the distance from the 3' CCA terminus to the target cytidine, specifically recognizing a UCCA motif (Long et al., 2016; Selmi et al., 2021). NSUN6 is also the specific writer for position C1218 of the Dengue virus genomic RNA, a modification essential for viral fitness (Wu et al., 2026; Zhang et al., 2025), which is used as a cross-RNA validation target in this thesis (Section 3.2). The remaining NSUN family members (NSUN3, NSUN4, NSUN7) and DNMT2, which serve distinct biological roles but are not directly investigated here, are summarized in Table 1.1.

Table 1.1 The m^{5}C Writer Enzymes.

CLASS I: RIBOSOMAL RNA METHYLTRANSFERASES

Enzyme	Organism	Substrate	Site	Function	Reference
Nop2	*S. cerevisiae*	25S rRNA	C2870	PTC-proximal; 60S biogenesis	Sharma et al., 2013
NSUN1/NOP2	Human	28S rRNA	C4447	Ribosome assembly	Liao et al., 2022
Rcm1	*S. cerevisiae*	25S rRNA	C2278	Stress translation; lifespan	Schosserer et al., 2015
NSUN5	Human	28S rRNA	C3782	Translational fidelity	Heissenberger et al., 2019
NSUN4	Human	12S mt-rRNA	C911	Mitoribosome; MTERF4 complex	Camara et al., 2011

CLASS II: TRANSFER RNA METHYLTRANSFERASES

Enzyme	Organism	Substrate	Site	Function	Reference
NSUN2	Human	Cytosolic tRNAs	C34, C48-50	Stability; Dubowitz syndrome	Blanco et al., 2014
NSUN2	Human	mt-tRNA^{Leu,Glu}	Variable loop	Mito translation	Van Haute et al., 2016
NSUN3	Human	mt-tRNA^{Met}	C34 → f^{5}C	Wobble; mito myopathy	Van Haute et al., 2016; Nakano et al., 2016
NSUN6	Human	tRNA^{Cys,Thr,Ser}	C72	Molecular ruler; 3'CCA	Haag et al., 2015; Long et al., 2016
DNMT2	Human	tRNA^{Asp,Gly,Val}	C38	Stress; hematopoiesis	Goll et al., 2006; Tuorto et al., 2015

CLASS III: MESSENGER RNA METHYLTRANSFERASES

Enzyme	Organism	Substrate	Site	Function	Reference
NSUN2	Human	mRNA	CDS, 3'UTR	Export (ALYREF); stability	Yang et al., 2017
NSUN6	Human	mRNA 3'UTR	CTCCA motif	High stoichiometry	Selmi et al., 2021; Zhang et al., 2025
NSUN7	Human	mRNA; eRNA	Variable	CCDC9B stability; liver cancer	Aguilo et al., 2016; Ortiz-Barahona et al., 2023

CLASS IV: RNA METHYLTRANSFERASES FOR NON-CODING AND VIRAL RNA

Enzyme	Context	Substrate	Site	Function	Reference
NSUN2	Human	Vault RNA	vtRNA1.1	Small RNA processing	Hussain et al., 2013
NSUN2	Human	lncRNA	XIST, HOTAIR	lncRNA function	Squires et al., 2012
NSUN6	DENV	Viral gRNA	C1218	Viral fitness	Wu et al., 2026

1.2.3.3 Readers and Erasers

The biological consequences of m^5C are mediated by reader proteins that recognize the methylated base. ALYREF mediates the export of m^5C-modified mRNAs from the nucleus to the cytoplasm by recruiting the nuclear pore machinery (Yang et al., 2017). In the cytoplasm, YBX1 binds to m^5C-modified transcripts, preventing their degradation; in bladder cancer models, this interaction drives the invasive phenotype (Chen et al., 2019). The dynamic nature of the epitranscriptome implies that m^5C can also be removed. The TET family of dioxygenases can oxidize m^5C to 5-hydroxymethylcytosine (hm^5C) in both tRNA and mRNA, suggesting that m^5C is part of a tunable regulatory circuit (Fu et al., 2014).

1.2.4 Stoichiometry and Biological Dynamics of m^5C

The biological significance of a modification is defined not merely by its presence but by its stoichiometry, the fraction of modified versus unmodified molecules at a specific site. Unlike the genome, which is largely static, the transcriptome exists as a dynamic population where individual RNA molecules can differ in their modification status. At any given genomic coordinate, a specific cytidine on a particular mRNA species may exist in a fractional state. Some copies carry the m^5C mark while others do not. This distribution is not random noise but represents a sophisticated layer of dosage control (Squires et al., 2012; Acera Mateos et al., 2024). These sub-populations often follow divergent fates; the methylated fraction might be preferentially recognized by reader proteins for nuclear export and translation, while the unmethylated fraction is retained or targeted for degradation (Yang et al., 2017).

Current evidence suggests that stoichiometry can be highly variable across sites and modification types (Garcia-Campos et al., 2019; Schumann et al., 2020). For m^6A, there is evidence that at least for a subset of sites, stoichiometry is “hard coded” in cis via sequence context. There remains ongoing debate about whether changes in stoichiometry and sub-stoichiometric sites have specific biological functions or are primarily caused by artifacts from experimental and computational detection methods. This uncertainty underscores the critical need for detection methods that can accurately quantify modification levels at single-molecule resolution, a central challenge that this thesis addresses through the development of deep learning-based approaches for nanopore DRS data.

1.2.5 2’-O-Methylation: Chemistry, Biology, and Relevance to Viral RNA

While 5-methylcytidine represents a nucleoside modification, 2'-O-methylation (Nm) targets the ribose sugar itself, converting the reactive 2'-OH to an inert 2'-OCH_3 (Helm & Motorin, 2017). This modification is among the most abundant in cellular RNA. rRNA alone contains over 100 Nm sites in higher eukaryotes, installed by snoRNA-guided machinery centered on the catalytic methyltransferase fibrillarin (Nop1 in yeast), and tRNA harbors several conserved Nm positions critical for structural integrity and translational fidelity (Sloan et al., 2017).

The chemical consequences of 2'-O-methylation are distinct from those of base modifications. By removing the nucleophilic 2'-hydroxyl group, Nm confers resistance to alkaline and enzymatic hydrolysis, a property forming the basis for its biochemical detection (Section 1.3.4). The modification also constrains the ribose into the C3'-endo conformation, increasing local backbone rigidity and enhancing base-stacking interactions (Kawai et al., 1992). In rRNA, these structural effects cluster at functionally important sites such as the peptidyl transferase center and decoding site, where they contribute to the precision of protein synthesis. Notably, the recent identification of sub-stoichiometric Nm sites in rRNA has overturned the assumption that all rRNA modifications are constitutively present, highlighting 2'-O-methylation as a potential source of ribosomal heterogeneity with regulatory implications (Sloan et al., 2017).

In the context of viral RNA biology, 2'-O-methylation plays a critical role in innate immune evasion. The mammalian innate immune system distinguishes self from non-self RNA partly through the recognition of 2'-O-methylation patterns. Unmodified RNA is detected as foreign by cytoplasmic sensors such as the IFIT family of proteins, which bind to RNA lacking 2'-O-methylation at the 5' cap (cap-1 structure) and restrict its translation (Daffis et al., 2010). Viral methyltransferases, including the SARS-CoV-2 Nsp16/Nsp10 complex, catalyze cap methylation to evade this restriction (Decroly et al., 2012; Viswanathan et al., 2020). Whether internal Nm sites on viral coding regions serve additional regulatory or immune-evasive functions remains an open question that motivated the experimental studies described in Section 3.1.

Nm within coding sequences also affects mRNA translation in a position-dependent manner. Using single-molecule and bulk kinetic approaches, Choi et al. (2018) demonstrated that Nm within codons sterically perturbs interactions of ribosomal monitoring bases (G530, A1492, and A1493) with cognate codon-anticodon helices, disrupting tRNA selection during translation elongation. Complementary work by Hoernes et al. (2016) showed that the magnitude and direction of translational effects depend on which position within the codon is modified, with

certain positions causing premature termination while others remain tolerated. These observations motivated the functional translation studies described in Section 3.1.2.

1.3 Conventional RNA Modification Detection Methods

The exploration of RNA modification landscapes has been historically defined by the analytical tools available. While methods developed for highly abundant structural RNAs proved effective, the transition to the dynamic and sparse mRNA epitranscriptome has exposed technical ceilings across multiple modification types. The following sections critically evaluate the principal approaches that have shaped m^5C and Nm research, highlighting their strengths while exposing the fundamental limitations that motivated the development of nanopore-based detection strategies.

1.3.1 LC-MS/MS (Liquid Chromatography-Mass Spectrometry)

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) remains the gold standard for qualitative and quantitative identification of modified ribonucleosides (Kellner et al., 2010; Thuring et al., 2016). The conventional workflow involves total enzymatic hydrolysis of RNA into constituent mononucleosides, followed by separation via high-performance liquid chromatography and analysis by mass spectrometry. By measuring mass-to-charge ratios and fragmentation patterns through collision-induced dissociation, LC-MS/MS provides absolute chemical specificity with sub-femtomole sensitivity (Basanta-Sanchez et al., 2016; Limbach et al., 1994). However, because RNA must be digested into single nucleosides, all positional and sequence information is lost (Kowalak et al., 1993; Crain & McCloskey, 1998). While the method confirms that m^5C or Nm exists within a sample and quantifies its total abundance, it cannot assign the modification to a specific transcript or genomic coordinate, rendering it a bulk tool unsuitable for site-specific analysis.

1.3.2 Bisulfite Sequencing (BS-seq)

RNA bisulfite sequencing (RNA-BS-seq) has been the most widely used approach to interrogate m^5C at single-nucleotide resolution, adapted from the bisulfite reaction originally developed for DNA methylation analysis (Frommer et al., 1992). RNA is treated with sodium bisulfite under acidic conditions, promoting deamination of cytosine to uracil, while m^5C is comparatively resistant to conversion (Schaefer et al., 2009). During reverse transcription and sequencing, converted cytosines are read as T, whereas positions retaining C are interpreted as candidate

m^5C sites. The readout is inherently quantitative. The residual C/(C+T) fraction at each position estimates methylation stoichiometry (Squires et al., 2012).

Despite its conceptual simplicity and base resolution, RNA-BS-seq is technically demanding. The same reaction conditions needed for high conversion efficiency impose harsh chemical stress that frequently leads to RNA fragmentation and loss of library complexity, reducing coverage uniformity and limiting sensitivity for low-abundance transcripts (Trixl & Lusser, 2019).

A second challenge is that incomplete conversion (non-conversion) of canonical cytosine can mimic methylation and represents a major source of false positives. This is particularly problematic for mRNA, where many reported sites are sub-stoichiometric. Distinguishing true low-occupancy m^5C signal from small non-conversion background demands high read depth, replicate consistency, and stringent filtering (Legrand et al., 2017). Non-conversion is not random; it is influenced by local RNA secondary structure and accessibility, meaning structured regions can exhibit elevated residual C even in the absence of m^5C (Trixl & Lusser, 2019).

Finally, bisulfite conversion introduces a substantial information and alignment penalty. Because bisulfite conversion changes nearly all cytosines to uracils, the resulting sequences consist predominantly of only three bases (A, G, and U/T), drastically reducing the information content available for unique alignment. This three-letter alphabet increases the probability that short reads map ambiguously to multiple genomic locations, particularly in repetitive elements and among highly similar transcript isoforms. Without careful bisulfite-aware alignment and conservative calling, mis-mapping can yield artifactual retained C signals that resemble methylation (Legrand et al., 2017). These constraints position RNA-BS-seq as a benchmark chemistry for nucleotide-resolution validation in targeted designs where depth and controls can be maximized, but have motivated the development of bisulfite-free alternatives for transcriptome-wide discovery (Lu et al., 2024).

1.3.3 Antibody-Based and Enzyme-Assisted Methods

m^5C-RNA Immunoprecipitation (RIP) sequencing uses anti-m^5C antibodies for immunoprecipitation followed by high-throughput sequencing. While comparatively gentle on RNA integrity, it typically yields peak-level enrichment (~100-200 nt resolution) rather than precise nucleotide calls. The primary limitation is dependence on antibody specificity and affinity, which are notoriously variable for small molecule epitopes. Off-target binding to abundant housekeeping transcripts can lead to spurious peaks, and RNA secondary structure may mask modification sites from antibody access (McIntyre et al., 2020).

Two enzyme-assisted strategies exploit the catalytic mechanism of m^5C methyltransferases to achieve single-nucleotide resolution. In miCLIP (methylation individual-nucleotide-resolution crosslinking and immunoprecipitation), the methyltransferase carries a single amino acid mutation (e.g., C271A in NSUN2) so that it binds to the RNA substrate but cannot be released, creating a covalent enzyme-RNA complex that can be immunoprecipitated (Hussain et al., 2013). In Aza-IP (5-azacytidine-mediated RNA immunoprecipitation), irreversible binding forms between the 5-azaC nucleoside analog and the RNA methyltransferase when exogenously provided 5-azaC is incorporated into nascent RNA (Khoddami & Cairns, 2013). While both methods achieve single-nucleotide resolution, they share significant limitations. Both require prior knowledge of the writer enzyme, are limited to cultured cells, and can potentially disturb the endogenous m^5C methylome through enzyme overexpression or analog toxicity (Lu et al., 2024). m^5C-TAC-seq (TET-Assisted Chemical labeling sequencing) employs TET-assisted oxidation to convert m^5C to 5-formylcytosine (f^5C), which is then selectively labeled through chemical derivatization, enabling enrichment and base-resolution profiling without the RNA degradation associated with bisulfite treatment (Lu et al., 2024; Zhang et al., 2025).

1.3.4 Detection Methods for 2'-O-Methylation

The detection of Nm requires fundamentally different analytical strategies than those used for base modifications, because ribose methylation does not alter base-pairing properties and is therefore invisible to bisulfite-based or antibody-based approaches. The available methods exploit three distinct biochemical properties of the 2'-O-methyl group. The first strategy leverages the capacity of Nm to stall reverse transcriptase under limiting dNTP (deoxynucleotide triphosphate) concentrations. At low dNTP, the polymerase pauses one nucleotide before a 2'-O-methylated site, producing truncated cDNA products that can be resolved by primer extension or by the high-throughput adaptation RTL-P (Reverse Transcription at Low dNTP concentrations followed by PCR) (Dong et al., 2012). The second strategy employs enzymatic selectivity: NJU-seq exploits the resistance of Nm to cleavage by Mycoplasma genitalium RNase R, and when combined with the site-specific quantification tool Nm-VAQ, recently revealed thousands of substoichiometric Nm sites on mRNA (Tang et al., 2024). The third strategy exploits differential susceptibility to chemical cleavage. Nm-seq and RibOxi-seq use periodate oxidation to selectively eliminate fragments with unmodified 2',3'-cis-diol termini while retaining those ending in 2'-O-methylated residues, enabling transcriptome-wide Nm mapping at base resolution (Dai et al., 2017; Zhu et al., 2017). RiboMethSeq, the approach most directly relevant to this thesis, instead measures the resistance of Nm sites to alkaline hydrolysis

(Birkedal et al., 2015; Marchand et al., 2016). Unlike the oxidative-selection methods, RiboMethSeq also provides quantitative measurement of absolute Nm stoichiometry, as ScoreC2 values show reasonable correlation with LC-MS/MS-derived methylation fractions (Motorin & Marchand, 2018) and is described in detail below.

RiboMethSeq exploits the resistance of Nm positions to alkaline hydrolysis. Under controlled alkaline conditions, unmodified ribonucleotides undergo intramolecular transesterification through nucleophilic attack of the 2'-hydroxyl on the adjacent phosphodiester bond, producing strand cleavage. Because 2'-O-methylation replaces the reactive hydroxyl with a methyl ether, methylated positions are refractory to this cleavage, producing characteristic gaps in fragment end coverage when reads are mapped by high-throughput sequencing (Marchand et al., 2016).

Quantification is based on the ScoreC2 metric, which measures the local depletion of fragment ends at a candidate position relative to flanking nucleotides. Values approaching 1.0 indicate strong protection consistent with complete methylation, while values near 0 indicate normal cleavage. Positions with ScoreC2 exceeding a defined threshold (typically 0.6) are called as candidate Nm sites. This approach has been successfully applied to map the complete Nm landscape of ribosomal RNA, where modification sites are dense and well-characterized (Marchand et al., 2016; Ayadi et al., 2018). In addition to ScoreC2, two complementary metrics are commonly employed. ScoreMEAN averages the protection scores over a broader window (typically ±6 nucleotides) to smooth local coverage fluctuations and is used as a high-confidence filter for candidate site selection. ScoreA (or ScoreA2) provides an alternative quantification based on the ratio of fragment ends at the candidate position to those at immediately adjacent nucleotides, offering a narrower but more position-specific measure of protection. In the analyses presented in this thesis, candidate sites were initially filtered using ScoreMEAN ≥ 0.92 and ScoreA2 ≥ 0.5, while ScoreC2 served as the primary metric for evaluating individual positions.

However, the reliability of threshold-based ScoreC2 calling has been critically examined. Pichot et al. (2022) applied a Random Forest machine learning algorithm to RiboMethSeq analysis, demonstrating that subjective threshold-based classification, while appropriate for limited-complexity datasets such as isolated rRNA, leads to substantial false-positive and false-negative identifications when applied to longer or more complex reference sequences. The fundamental limitation is that any factor reducing cleavage efficiency such as stable RNA secondary structure, sequence-dependent cleavage bias, or ligation junction artifacts, produces protection signals indistinguishable from genuine Nm (Helm & Motorin, 2017; Motorin & Marchand, 2018). Highly structured RNA regions require substantially more sequencing reads than theoretically

predicted, and the ScoreC2 metric can produce artifacts in regions of irregular fragmentation (Motorin & Marchand, 2018).

1.3.5 Fundamental Limitations of Conventional Methods

The methodologies described above have defined much of what is currently known about the m^5C and Nm landscapes, yet they share a fundamental constraint. They are primarily ensemble measurements. Whether relying on chemical conversion (BS-seq), antibody enrichment (m^5C-RIP), or differential hydrolysis (RiboMethSeq), these workflows integrate signals across large populations of molecules, yielding an average that can mask biologically meaningful heterogeneity (Acera Mateos et al., 2023).

This ensemble-averaging bottleneck restricts at least two central questions. First, it obscures stoichiometric heterogeneity. An apparent 50% methylation level at two nearby sites is compatible with very different molecular realities (e.g., half the molecules modified at both sites versus mutually exclusive modification across molecules) (Squires et al., 2012). Second, it prevents the study of modification phasing, whether two modifications on the same transcript tend to co-occur on the same molecule or are mutually exclusive across different copies. This distinction is critical because combinatorial modification patterns may jointly determine RNA fate, yet ensemble methods cannot distinguish these scenarios (Acera Mateos et al., 2024).

Furthermore, indirect biochemical readouts are inherently susceptible to context-dependent artifacts that can mimic genuine modification. The common thread is that each method's signal depends on an intermediate chemical or enzymatic step whose efficiency varies with RNA structure, sequence, and preparation conditions (Trixl & Lusser, 2019; Legrand et al., 2017). These limitations have created a demand for technologies capable of reading modifications on single, native RNA molecules without chemical conversion or amplification. This technological gap has been addressed by the emergence of nanopore DRS, which offers an orthogonal detection paradigm by reading ionic current signatures from individual molecules as they translocate through protein nanopores (Garalde et al., 2018).

1.4 Nanopore Direct RNA Sequencing (DRS)

Nanopore sequencing has introduced a fundamentally different measurement modality for transcriptomics. Instead of inferring sequence from optical signals generated during sequencing-by-synthesis, Oxford Nanopore Technologies (ONT) instruments measure ionic

current through a nanoscale biological pore while a single RNA strand is threaded through it under an applied electric field (Figure 1.4; Deamer et al., 2016). This real-time electrical readout enables single-molecule observations with long read lengths that preserve transcript context (Garalde et al., 2018). For epitranscriptomics, the measured signal originates from the native RNA strand rather than an amplified DNA copy, maintaining molecule-resolved information lost in short-read workflows (Garalde et al., 2018; Fleming & Burrows, 2026). Importantly, direct RNA should not be conflated with "no enzymes". Current ONT DRS library preparation includes one round of reverse transcription to form a DNA:RNA duplex used by the motor system, but the platform's defining feature is that it senses the RNA strand electrically without requiring cDNA amplification (Fleming & Burrows, 2026).

1.4.1 Physical Principles and Fundamental Terminology

The conceptual basis of nanopore sensing is analogous to the Coulter counter principle, where passage of an analyte through a constriction modulates an electrical signal. A protein nanopore is embedded in an electrically resistant membrane separating two electrolyte-filled chambers. When voltage is applied, ions flow through the pore, producing an open pore current, a baseline ionic current in the absence of a strand (Kasianowicz et al., 1996; Deamer et al., 2016).

As a polyanionic RNA strand enters and traverses the pore under the electrophoretic force, it partially occludes the pore lumen and perturbs ion flow, producing an ionic current blockade recorded in the picoampere (pA) range (Kasianowicz et al., 1996). A motor protein ratchets the nucleic acid through the pore at a controlled pace, setting the translocation rate and shaping the distribution of dwell times, the time a given sequence context spends in the pore's constriction zone (Derrington et al., 2015; Fleming & Burrows, 2026). Because motor kinetics depend on duplex stability and local chemistry, motor-associated features, especially dwell time, can sometimes carry modification information in addition to current amplitude (Fleming et al., 2021; Fleming et al., 2023). The resulting time series of current values is called the raw signal or "squiggle," a characteristic step-wise trace in which each plateau corresponds to a distinct current level produced by a particular k-mer occupying the pore constriction. The squiggle thus encodes both sequence and chemical identity. Any covalent modification within the sensed k-mer can shift the current level, alter the dwell time, or both, relative to the unmodified reference.

A key concept for modification detection is that the ionic current at any moment reflects a multi-nucleotide context within the pore constriction (Figure 1.4, lower right). For Nanopore DRS, the effective context is often described as 5 nucleotides; however, under the recent chemistry RNA004, ONT's released k-mer model has been reported as a 9-mer model, with evidence that

the central bases carry much of the discriminative contribution (Makhamreh et al., 2024; Samarakoon et al., 2025). Modification signatures are therefore context-dependent perturbations distributed across the local k-mer neighborhood rather than single-base events. This has two major consequences: (i) the same modified base can produce different signal perturbations depending on its surrounding k-mer (context-dependence), and (ii) accurate calling typically requires computational models that map squiggle segments to sequence and modification states rather than relying on a single current value (deconvolution requirement). These constraints become particularly important for m^{5}C, whose physicochemical perturbation is often subtle relative to stronger-shifting modifications (Fleming & Burrows, 2026).

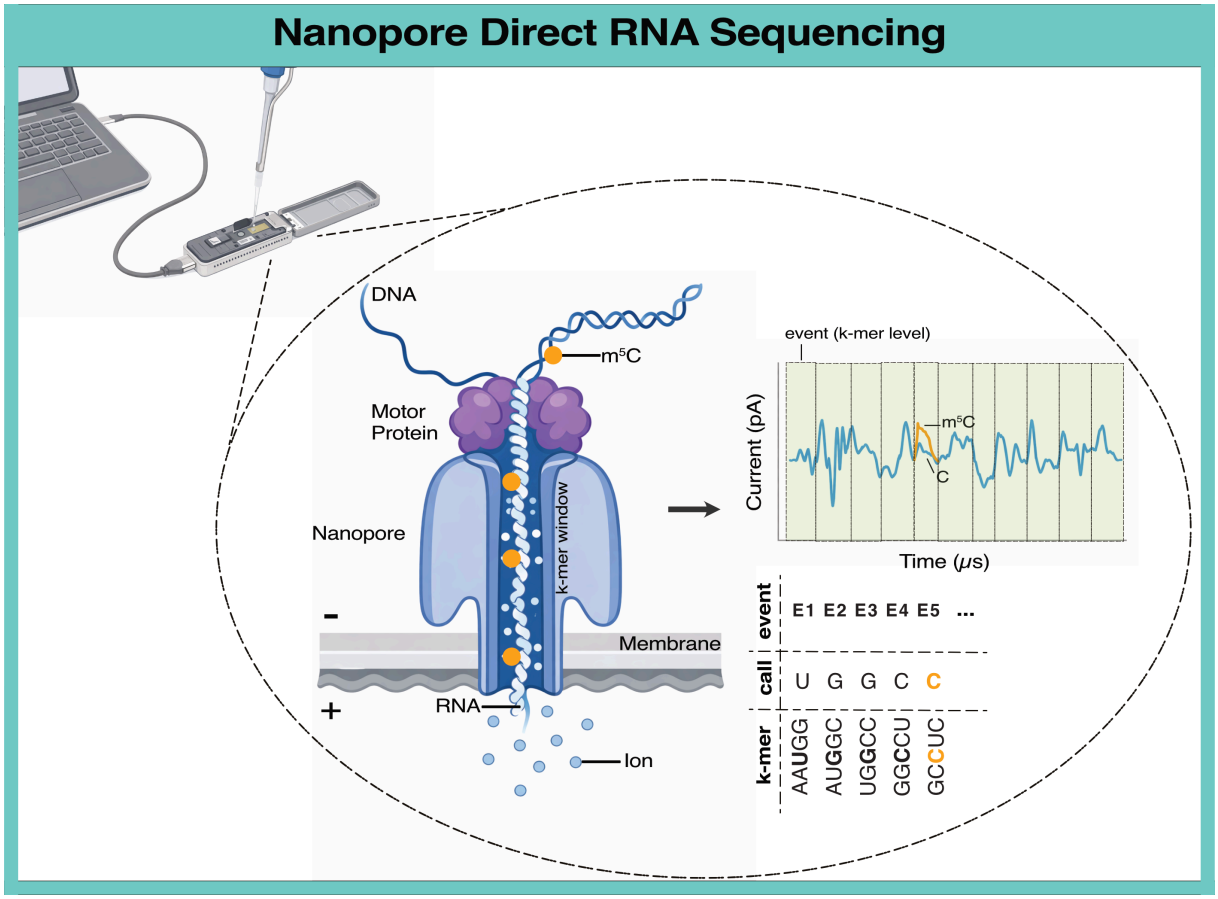


Figure 1.4. Nanopore direct RNA sequencing principle and modification detection. A single RNA strand (white) hybridized to a DNA adapter (blue) is ratcheted through a protein nanopore embedded in a synthetic membrane by a motor protein. As the RNA translocates, it modulates the ionic current (pA), producing a raw signal trace (upper right). Each discrete current level corresponds to an event determined by the k-mer sequence occupying the pore constriction at that moment. The basecaller converts events (E1–E5) into nucleotide calls, with each event reflecting a 5-mer context (lower right). A chemically modified nucleotide such as m^{5}C (orange) alters the current profile relative to unmodified cytidine (C), providing the basis for modification detection. The MinION sequencing device connected to a laptop is shown at upper left. MinION device and computer illustrations adapted from Oxford Nanopore Technologies. Nanopore protein illustration adapted from BioRender (both accessed 30 January 2025). All other elements created by the author.

1.4.2 Modification Detection via Nanopore DRS

Raw current traces are converted to nucleotide sequence through basecalling. Early basecalling relied on Hidden Markov Models with event segmentation and k-mer current models (David et al., 2017), while later approaches adopted deep neural networks that learn complex signal-to-sequence mappings directly from data, substantially improving accuracy (Boza et al., 2017). ONT's primary basecaller has evolved from Guppy to Dorado (Fleming & Burrows, 2026). Many modification-detection workflows additionally require signal-to-sequence alignment, a process known as resquiggling. During resquiggling, the raw ionic current trace is computationally re-aligned to the reference genome sequence so that each segment of the signal is assigned to a specific nucleotide position. This is achieved by fitting the observed current levels to an expected k-mer model using dynamic programming or segmentation algorithms, effectively reversing the basecalling abstraction and recovering per-position signal features such as current mean, standard deviation, and dwell time. Resquiggling is essential for position-level comparison between conditions and for extracting the quantitative features that downstream modification-detection tools use as input (Stoiber et al., 2017). Practically, the feasibility and comparability of modification detection are shaped by rapid evolution of ONT chemistries. Shifts in pore/motor/basecaller versions can measurably change signal properties and alter modification-calling behavior across runs unless versions are carefully controlled (Samarakoon et al., 2025).

The distinction between squiggling, basecalling, sequence alignment, and resquiggling is illustrated in Figure 1.5. Basecalling converts the raw squiggle into a nucleotide sequence (FASTQ) but discards the underlying signal information. Subsequent sequence alignment (e.g., via minimap2) maps this read to the reference genome in sequence space, producing a BAM file, yet still retains no signal-level data. Resquiggling closes this gap by re-aligning the original ionic current trace to reference coordinates using dynamic time warping (DTW; Tombo; Stoiber et al., 2017) or hidden Markov models (HMM; Nanopolish; Simpson et al., 2017), thereby recovering per-position features, including current mean, dwell time, and signal variance, that are indispensable for downstream modification detection.

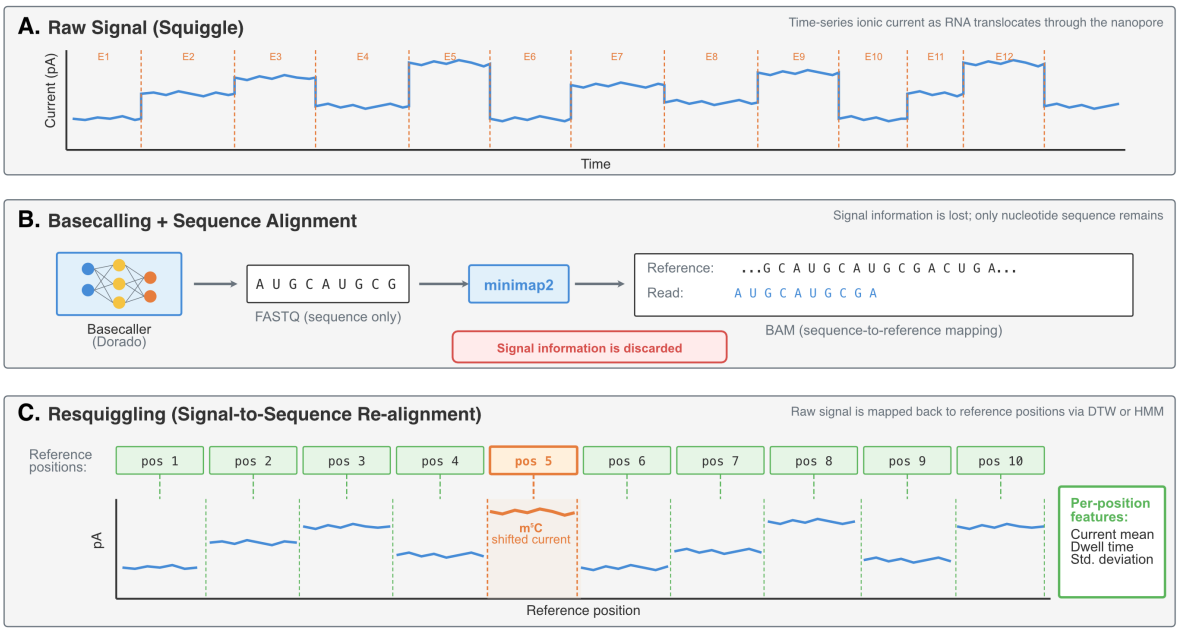


Figure 1.5. From squiggle to per-position features: basecalling, alignment, and resquiggling. (A) The raw signal (squiggle) recorded during nanopore DRS is a step-wise ionic current time series, with each event (E1–E12) corresponding to a distinct k-mer occupying the pore constriction. **(B)** Basecalling by a neural network (e.g., Dorado) converts the squiggle into a nucleotide sequence (FASTQ), which is then aligned to a reference genome by minimap2 (BAM). Critically, signal-level information is discarded during this process. **(C)** Resquiggling re-aligns the original raw signal to reference positions using DTW or HMM algorithms guided by a k-mer pore model, recovering per-position features (current mean, dwell time, standard deviation) required for modification detection. The highlighted position (orange, pos 5) illustrates how a chemical modification such as m^5C shifts the ionic current relative to canonical expectation.

1.4.3 RNA Modification Detection Tools Using Nanopore DRS

A natural next step after defining the DRS measurement pipeline is to ask how the field first extracted modification information from this signal. RNA modification detection was largely pursued through three complementary computational paradigms: basecalling-error frameworks, signal-level comparative frameworks, and feature-engineered machine learning classifiers.

1.4.3.1 Basecalling-Error Frameworks

Early studies established that RNA modifications can induce reproducible basecalling perturbations such as mismatches, indels, and lower base qualities, because canonical basecallers are trained largely on unmodified sequence and therefore interpret modified k-mers imperfectly. A landmark demonstration showed that m^6A can be detected from systematic basecalling errors and quality shifts in ONT DRS, providing a general proof-of-principle that error profiles carry modification information (Liu et al., 2019). This observation motivated a family of tools that quantify per-position error signatures and compare them between

conditions. ELIGOS formalized this via a per-base "error of specific bases" concept and differential testing (Jenjaroenpun et al., 2021). EpiNano calculated summary statistics of basecalling accuracy, mismatch frequency, and deletion rates at each genomic position, then used machine learning to classify m^6A (Liu et al., 2021). DRUMMER used similar principles with support for replicates (Abebe et al., 2022), and JACUSA2 extended multivariate error modeling to multiple modification types in targeted contexts (Naarmann-de Vries et al., 2023). A fundamental limitation is that error-based signatures are indirect. They depend on how a modification confuses the basecaller rather than measuring it directly. This indirect readout can be confounded by sequence context, secondary structure, and basecaller version. For m^5C specifically, error patterns have been reported to be inconsistent enough to limit confident identification (Jenjaroenpun et al., 2021).

1.4.3.2 Signal-Level Comparative Frameworks

Rather than relying on basecalling error, signal-level approaches directly compare raw current distributions at each position between conditions. A key enabling step is signal-to-sequence alignment (resquiggling), which maps the time series onto nucleotide positions so that per-site comparisons are meaningful. A foundational toolkit for this paradigm is Tombo, which popularized genome-guided signal processing and comparative testing workflows (Stoiber et al., 2017).

Building on this strategy, Nanocompore (Leger et al., 2021) introduced a replicate-aware comparative framework that tests whether per-position signal distributions differ between a sample and a control, demonstrating detection across multiple modification types. xPore framed the problem as inference of differential modification rate between conditions using a probabilistic mixture model, enabling detection of positions whose modification fraction shifts across samples, an especially useful framing for dynamic regulation and perturbation designs (Pratanwanich et al., 2021). Yanocomp similarly leveraged mixture modeling with support for replicates and single-molecule-level assignments (Parker et al., 2021). Compared to error-based methods, signal-level approaches are conceptually closer to the underlying physics; however, they still face context dependence, run-to-run variability, and strong dependence on well-matched controls (IVT, enzymatic depletion, or writer perturbations). These constraints become particularly consequential for subtle marks such as m^5C, where effect sizes can be small relative to baseline signal variance (Leger et al., 2021; Pratanwanich et al., 2021).

1.4.3.3 Feature-Engineered Machine Learning Classifiers

A third class of tools introduced conventional supervised classifiers to improve upon pure thresholding or single-test statistics. Nanopore reads are first featurized into compact descriptors, typically summary statistics of the aligned current signal and timing behavior, and then classified using standard machine learning models such as Random Forest classifiers trained on experimentally supported sites. Prominent examples include EpiNano (Liu et al., 2021) and MINES (Lorenz et al., 2020). Although effective within their designed scope, these classifiers share two practical limitations. Their performance is tightly coupled to the feature-engineering choices and training labels, so domain shift (new species, chemistry, or modification context) often requires re-training; and feature compression can discard signal information that a richer model might exploit, limiting sensitivity for subtle or low-stoichiometry modifications such as m^5C.

1.4.4 Limitations of Classical Nanopore Approaches and the Transition to Deep Learning

The tools described above represent important milestones, yet they share structural constraints that become limiting as epitranscriptomic questions grow more demanding. Error-based methods depend on basecaller behavior, which changes across software releases and chemistries; they are inherently indirect and have been shown to be unreliable for m^5C (Jenjaroenpun et al., 2021). Comparative/statistical methods require matched controls, often knockout or IVT references, adding experimental complexity and cost; they also report population-level differential metrics rather than per-read probabilities, limiting single-molecule resolution (Leger et al., 2021). Feature-engineered ML classifiers compress signal information and therefore may miss context-specific patterns that a richer representation could capture.

The transition toward deep learning is therefore best viewed not as eliminating these constraints, but as changing where they act and how strongly they limit performance. Deep models can learn richer, context-aware representations from raw or minimally processed signal, reducing reliance on brittle, hand-crafted features and improving discrimination for subtle marks such as m^5C. At the same time, the field still depends on ground truth design and controls (often IVT or writer perturbations) for training, calibration, and false-positive auditing, and many practical pipelines still require signal-to-reference mapping. The key shift is that deep learning enables single-sample inference at scale once trained, while classical toolkits often require explicit contrastive designs for each new dataset (Acera Mateos et al., 2024). This motivation provides the foundation for the deep learning strategies discussed in the following section.

1.5 Deep Learning in Nanopore Direct RNA Sequencing

1.5.1 Fundamentals of Deep Learning

1.5.1.1 Core Concepts and Terminology

Deep learning (DL) is a subfield of machine learning that uses multi-layered artificial neural networks to automatically learn feature representations from data. Each layer consists of neurons that apply weighted transformations and nonlinear activation functions (Figure 1.6); stacking many layers enables the model to learn complex, hierarchical patterns (Goodfellow et al., 2016). During training, network weights are iteratively adjusted via stochastic gradient descent to minimize a loss function over multiple epochs (complete passes through the training data). Unlike classical machine learning algorithms (e.g., random forests, support vector machines) that require manual feature engineering, deep learning performs end-to-end learning, automatically extracting relevant features from high-dimensional data such as nanopore signals (Ching et al., 2018; Acera Mateos, 2023).

A key advantage of deep learning is that neural network architectures can be specifically tuned to suit assumptions about the data (inductive bias). Convolutional neural networks (CNNs) apply learnable filters (kernels) that slide across input data to detect local patterns. In the first CNN layer, kernels scan through all positions of the input vector and calculate an activation value for each position; the more relevant a particular feature combination, the higher the activation values. This makes CNNs well-suited for extracting shift-invariant features from sequential data such as nanopore current traces (LeCun et al., 1995). Recurrent neural networks (RNNs) model temporal dependencies by maintaining hidden states across time steps; however, standard RNNs suffer from vanishing gradient problems when learning long-range dependencies. Long Short-Term Memory (LSTM) networks address this through gating mechanisms that control information flow over extended sequences, enabling the network to learn which information to retain or discard (Hochreiter & Schmidhuber, 1997). While deep learning can outperform classical approaches, it tends to overfit in data-limited settings and typically requires GPU-accelerated computing (Salman & Liu, 2019; Pandey et al., 2022).

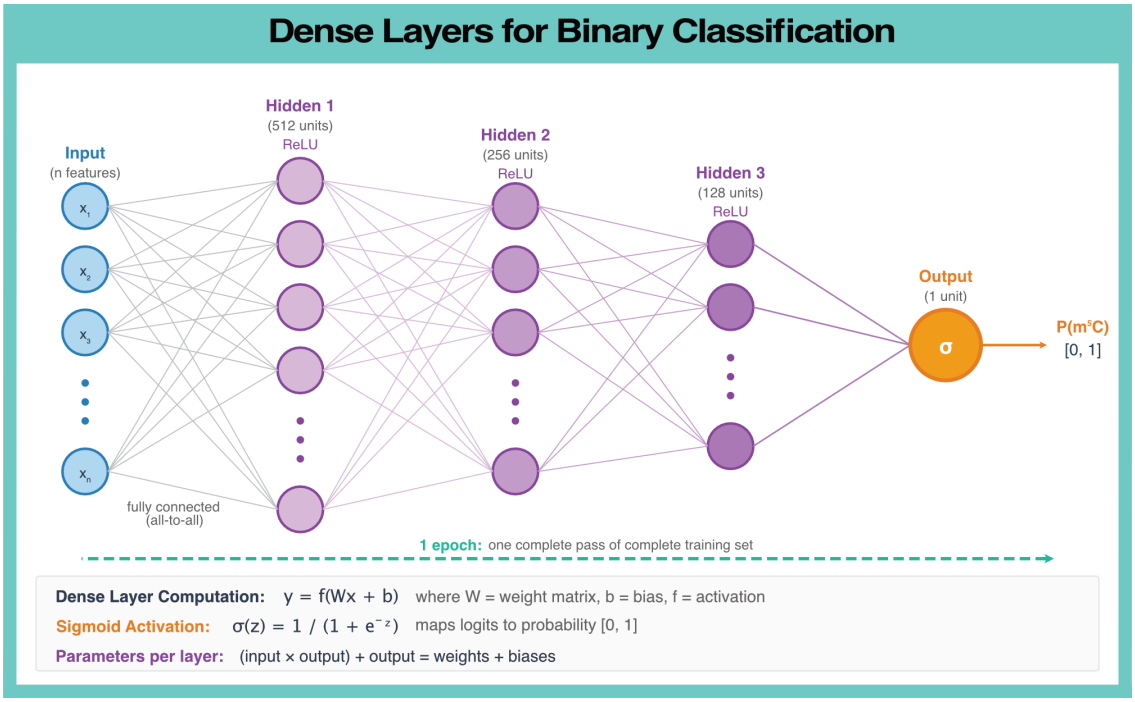


Figure 1.6. Dense layers for binary classification in RNA modification detection. A fully connected feed-forward neural network maps n input features (blue, x_1–x_n) through three hidden layers with decreasing dimensionality (512, 256, and 128 units; purple) using ReLU activation functions to a single output unit with sigmoid activation (σ, orange), producing a probability P(m⁵C) in the range [0, 1]. All neurons in adjacent layers are connected (all-to-all), with each connection carrying a learned weight. During training, the network iteratively adjusts weights over multiple epochs (one complete pass through the training data per epoch) to minimize a loss function. Key mathematical relations are shown below: the dense layer computation $y = f(Wx + b)$, the sigmoid function $\sigma(z) = 1/(1 + e^{-z})$, and the parameter count per layer (Goodfellow et al., 2016).

1.5.1.2 Evaluation Metrics for Classification Models

In binary classification tasks such as RNA modification detection, each prediction falls into one of four categories. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Sensitivity (recall, TP/(TP+FN)) measures detection of true modifications; specificity (TN/(TN+FP)) measures correct rejection of negatives; precision (TP/(TP+FP)) measures the fraction of positive predictions that are correct. The F1 score provides a balanced single metric as the harmonic mean of precision and recall (Sokolova & Lapalme, 2009). All of these metrics are scored on a scale from 0 to 1, with values approaching 1 indicating better performance.

The receiver operating characteristic (ROC) curve evaluates classifier performance across all possible thresholds by plotting sensitivity against false positive rate (Figure 1.7A). The area under the ROC curve (AUC) summarizes this as a single scalar, interpretable as the probability that a randomly chosen positive instance is ranked above a randomly chosen negative (Hanley & McNeil, 1982; Fawcett, 2006). An AUC of 0.5 corresponds to random discrimination (the diagonal in Figure 1.7A), while a value of 1.0 represents perfect classification; in practice, values

of 0.70–0.80 indicate acceptable, 0.80–0.90 excellent, and above 0.90 outstanding discrimination (Hosmer et al., 2013).

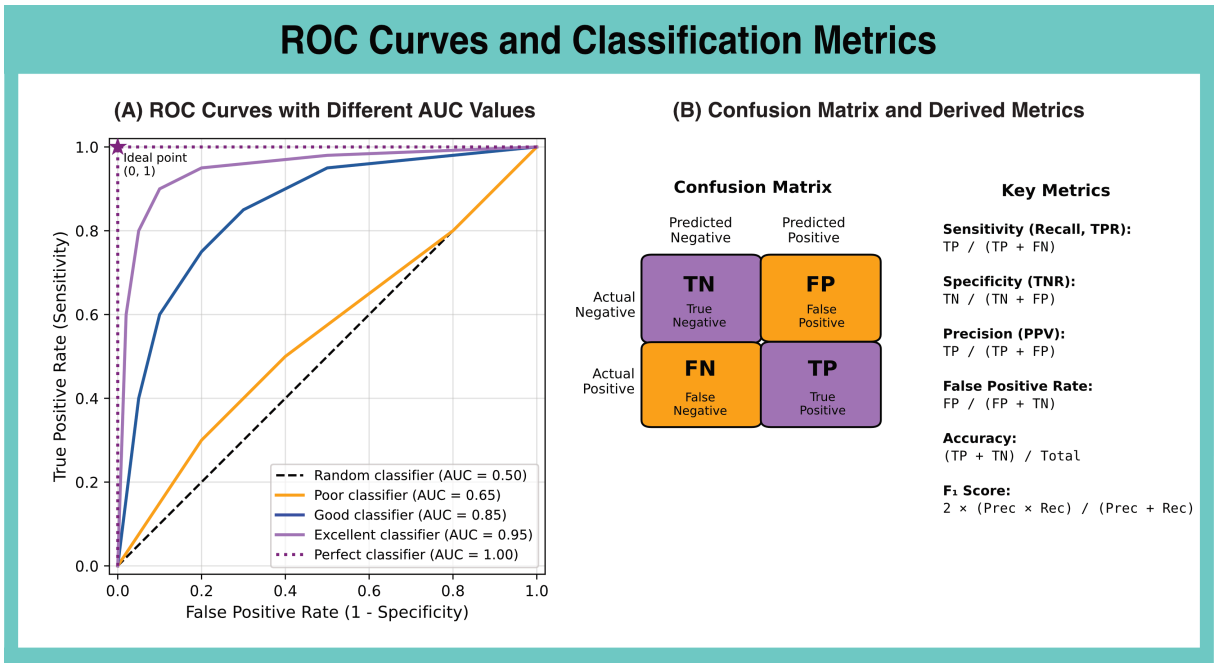


Figure 1.7. Receiver operating characteristic (ROC) curves and classification metrics. (A) ROC curves for classifiers with different discriminative abilities, ranging from random (AUC = 0.50, diagonal) to perfect classification (AUC = 1.00). The ideal point (0, 1) represents perfect sensitivity at zero false positive rate. **(B)** Confusion matrix showing the four prediction outcomes (TP, TN, FP, FN) and key derived metrics including sensitivity (recall), specificity, precision, false positive rate, accuracy, and F1 score.

Importantly, ROC-AUC has limitations for highly imbalanced datasets. In RNA modification detection, where unmodified positions vastly outnumber modified ones, even a seemingly low false positive rate can generate thousands of erroneous calls at transcriptome scale. Two complementary strategies address this imbalance. The first and more conservative approach is to balance the training data itself by equalizing the number of modified and unmodified examples presented to the classifier, ensuring that the model does not learn to default to the majority class. The second is to employ evaluation metrics that are sensitive to class distribution. The precision-recall curve and its area (AUPRC) are more informative than ROC-AUC for imbalanced problems because they focus on the positive (modified) class and expose the cost of false positives directly (Saito & Rehmsmeier, 2015).

1.5.2 Neural Network Architectures for Nanopore Signal Analysis

Several specialized neural network architectures have been applied to nanopore sequencing data, each exploiting different properties of the signal. The choice of architecture reflects assumptions about what patterns in the data are most informative for the task at hand.

1.5.2.1 Convolutional Neural Networks

The Jasper architecture, originally developed for speech recognition, exemplifies a deep CNN design that has been adapted for nanopore signal analysis (Li et al., 2019). CHEUI employs a modified Jasper CNN architecture that takes as input the signal values from five consecutive 5-mers (forming a 9-mer window) along with the distance between observed and expected signal values from an unmodified k-mer model; this distance metric improved classification accuracy by approximately 10%, demonstrating that combining raw signal features with derived features can enhance discrimination (Acera Mateos et al., 2024).

1.5.2.2 Recurrent Neural Networks and Long Short-Term Memory Networks

Early nanopore basecallers such as Chiron and DeepNano employed RNN and LSTM architectures to translate raw current signals into nucleotide sequences (Boza et al., 2017). MultiRM, designed to simultaneously detect twelve RNA modification types, combined LSTM layers with attention mechanisms to process input sequences and map them to multiple context vectors corresponding to each modification output (Song et al., 2021).

1.5.2.3 Transformer Architectures

Transformers represent a paradigm shift in sequence modeling by replacing recurrence with self-attention mechanisms that allow the model to weigh the importance of different positions in the input sequence when producing an output (Vaswani et al., 2017). This architecture enables parallel processing and can capture long-range dependencies more effectively than RNNs. DeepRM, a transformer-based model trained on a massive-scale dataset (over 297 million signal blocks totaling 6.24 billion nucleotides), achieved near-perfect accuracy in m^6A detection (AU-ROC = 0.997, AU-PR = 0.943) by using a 21-nucleotide context window and processing raw electric currents directly (Zhang et al., 2025). This success underscores that both training data scale and context length are critical determinants of model performance.

1.5.2.4 Hybrid and Specialized Architectures

Several tools combine multiple architectural elements for specific challenges. RedNano employs a hybrid framework combining raw signal features with basecalling-error features through residual networks, reflecting the pragmatic observation that signal-level and error-level cues can be complementary (Ni et al., 2024). m6Anet formulates site-level methylation as a multiple instance learning (MIL) problem using a feed-forward neural network, producing both site-level and read-level probabilities (Hendra et al., 2022). The MIL framework addresses the challenge that experimental methods identify modified sites at the population level but do not indicate

which individual molecules carry the modification; by treating each RNA site as a bag of reads labeled positive if the site is methylated, the network learns to infer which reads are modified without explicit read-level labels. Xron takes an end-to-end basecaller-like approach for m^6A detection, using a two-stage training strategy combining synthetic IVT-style supervision with immunoprecipitation-derived labels (Teng et al., 2024).

1.5.3 Deep Learning-Powered Tools for RNA Modification Detection

Deep learning has shifted nanopore DRS modification analysis toward models that learn modification-associated signal patterns directly, enabling per-read probabilities and more explicit stoichiometry estimation. In practice, DL-based DRS callers fall into three workflow-relevant classes. Single-modification frameworks, multi-modification frameworks, and modification-aware basecallers. A cross-cutting constraint across all three classes is chemistry dependence: RNA002 and RNA004 operate in different pore/motor regimes with altered signal statistics and effective sensing context, so model portability across kits should not be assumed without explicit chemistry-matched training or benchmarking (Zou et al., 2025; Zhang et al., 2025).

1.5.3.1 Single-Modification Frameworks

Most single-modification deep learning callers have focused on m^6A, which produces a relatively strong signal perturbation compared to other modifications. m6Anet remains a widely used reference model (Hendra et al., 2022); however, RNA004-era comparisons emphasize that calibration and absolute performance can shift across datasets and experimental regimes (Zou et al., 2025). SingleMod trains a multiple-instance regression framework using quantitative NGS-derived methylation-rate labels and is designed to resolve read-level heterogeneity (Xie et al., 2026). m6ATM uses a WaveNet encoder with a dual-stream MIL design and explicitly targets stoichiometry-oriented outputs validated on IVT mixtures and human data (Yu et al., 2024).

1.5.3.2 Multi-Modification Deep Learning Frameworks

A key challenge in nanopore DRS is that multiple modifications may fall within the pore's effective sensing context (typically 5–9 nucleotides), meaning their signals can interfere with one another. Multi-modification frameworks attempt to address this cross-talk problem while enabling simultaneous detection of several marks. CHEUI is a widely adopted dual-modification framework that predicts m^6A and m^5C in the same RNA002 sample, enabling single-molecule co-occurrence analyses across the transcriptome (Acera Mateos et al., 2024). NanoSPA extends

dual-modification detection to m^6A and pseudouridine, revealing opposing co-occurrence patterns and their synergistic effects on translation (Huang et al., 2024). TandemMod broadens the scope further by employing transfer learning to detect up to seven modification types, including m^6A, m^5C, m^1A, hm^5C, m^7G, inosine, and pseudouridine, within a unified RNA002 framework (Wu et al., 2024). modCnet addresses a cytidine-centered niche by training a deep learning model for simultaneous ac^4C and m^5C detection from RNA002 nanopore DRS using *in vitro*-constructed datasets (Wu et al., 2026). More recently, DirectRM introduced a two-stage pipeline combining binary detection with attention-based multi-label classification for six modifications (m^6A, m^5C, m^1A, m^7G, ac^4C, and pseudouridine), validated on both RNA002 and RNA004 chemistries (Zhang et al., 2025). ORCA takes a fundamentally different approach by employing domain adversarial learning rather than supervised classification, enabling *de novo* detection of modification types from RNA002 data without requiring labelled IVT training data; however, this unsupervised strategy trades the per-site quantitative calibration that supervised IVT-trained classifiers can provide (Dong et al., 2026). ModiDeC provides a multi-class inception–LSTM classifier for five RNA modifications and is accompanied by synthetic training datasets spanning both RNA002 and RNA004, reinforcing that chemistry shifts must be treated as model-relevant covariates (Alagna et al., 2025). Of particular relevance to the present work, ModiDeC serves as the base classifier upon which the calibration framework developed in this thesis is built (Section 3.2).

1.5.3.3 Modification-Aware Basecallers

A distinct and increasingly influential class comprises modification-aware basecallers, which integrate modified-base prediction into the basecalling process and output per-read, per-position modification scores as part of the standard analysis pipeline. ONT's Dorado basecaller supports detection of pseudouridine, m^6A, inosine, and m^5C, at the time of this study representing the only available m^5C caller compatible with the RNA004 chemistry, although DirectRM has since provided RNA004-compatible m^5C models whose independent validation remains limited (Zhang et al., 2025). The practical attraction of Dorado-style calling is that modified-base probabilities are emitted directly with the basecaller output and integrate seamlessly into existing workflows without requiring separate modification-calling steps.

1.5.4 The Challenge of m^5C Detection by Nanopore DRS

Among RNA modifications amenable to nanopore detection, m^5C presents challenges that have emerged consistently across independent benchmarking efforts. A large-scale comparative

evaluation identified m^5C as one of the poorest-performing modifications across multiple algorithms and both RNA002 and RNA004 chemistries, even when models were retrained on chemistry-matched data (Zou et al., 2025). This consistent underperformance reflects fundamental properties of the modification rather than implementation failures of any single tool. Several factors contribute. The physicochemical perturbation induced by m^5C produces subtle current shifts that are highly context-dependent; the typically low modification fraction at m^5C sites means that even modest false-positive rates can dominate overall positive predictions; and the scarcity of well-characterized, high-stoichiometry m^5C sites for training and validation limits model generalizability.

These challenges manifest across different tool categories. For modification-aware basecallers, independent benchmarking of Dorado revealed that up to 40% of predicted m^5C sites showed higher modification frequencies in unmodified IVT controls than in corresponding wild-type samples (Zou et al., 2025). In *Escherichia coli*, only one of three biochemically validated rRNA m^5C sites was detected, leading investigators to discontinue downstream analyses using the Dorado m^5C model (Cruciani et al., 2025). For multi-modification frameworks, CHEUI developers reported trade-offs between recall and false-positive rate for m^5C, along with limited overlap with bisulfite sequencing datasets (Acera Mateos et al., 2024). Even cytidine-specialized frameworks like modCnet explicitly treat m^5C as a confounding-prone regime requiring chemistry-aware modeling and careful orthogonal validation (Wu et al., 2026). These consistent challenges have driven the field to reconsider not only model architectures but also the training strategies required to reliably capture m^5C while minimizing false-positive predictions.

1.5.5 Training Strategies and Ground Truth Generation

1.5.5.1 Ground Truth Generation and Labeling Strategies

A central challenge in applying supervised deep learning to RNA modification detection is obtaining ground truth labels. Experimental methods can identify modified sites at the site level but do not usually indicate which individual RNA molecules carry the modification, meaning read-level labels are often missing. Multiple instance learning (MIL) addresses this by treating each RNA site as a bag of reads labeled positive if the site is methylated; the neural network learns to infer which reads are modified without explicit read labels (Hendra et al., 2022).

In vitro transcription (IVT) provides a controlled source of training data by producing RNA with known modification status. Fully modified IVT (where every instance of a nucleotide is replaced by its modified form) provides positive examples, while canonical IVT provides negative

examples. However, this uniformly modified regime differs substantially from the substoichiometric modification found in native transcripts. Splint ligation offers a refinement by enabling site-specific incorporation of a single modified nucleotide at a defined position within a precisely controlled sequence context, permitting training on modification signals unconfounded by neighboring modifications and providing ground truth at single-molecule resolution as exemplified by ModiDeC (Alagna et al., 2025). DeepRM employed chemical oligonucleotide synthesis to generate a massive dataset where each m^6A nucleotide was flanked by unmodified nucleotides, yielding detection accuracy far exceeding models trained on bulk IVT or transcriptome-derived data (Zhang et al., 2025).

Writer perturbation experiments (e.g., methyltransferase knockouts) provide biologically matched samples differing primarily in modification status at sites catalyzed by the targeted enzyme. These paired datasets serve both as supervision for model training and as validation for model performance under native conditions. However, secondary effects of writer perturbation and incomplete elimination of modification can introduce uncertainty into the ground truth.

A fundamental tension exists between synthetic ground truth (clean labels, limited context diversity) and native ground truth (biological relevance, uncertain labels). Synthetic constructs provide defined modification status at single-molecule resolution but lack the sequence diversity, RNA secondary structures, and neighboring modifications present in native transcripts. Models trained exclusively on synthetic data may learn decision boundaries that do not generalize to heterogeneous biological contexts. Conversely, native ground truth captures biologically relevant complexity but suffers from incomplete modification depletion, secondary cellular effects, and absence of read-level labels. Effective training strategies must navigate this trade-off, often by combining both data sources within carefully designed workflows.

1.5.5.2 Transfer Learning, Data Augmentation, and Class Imbalance

Transfer learning leverages knowledge from large source datasets to improve performance on related target tasks. In epitranscriptomics, a model pre-trained on the high-signal intensity of m^6A can be fine-tuned for m^5C: since the pore physics and signal noise are shared, fine-tuning the final layers on m^5C-specific labels enables the model to resolve subtle signatures without requiring massive amounts of new training data (Wu et al., 2024). TandemMod exemplifies this approach, expanding detection across multiple modification types within a unified framework.

Data augmentation strategies enhance model robustness by artificially expanding training set diversity. For nanopore signals, augmentation techniques include injecting Gaussian noise to simulate measurement variability, applying time warping to account for translocation speed

fluctuations, and signal scaling to mimic pore-to-pore differences. These transformations encourage the network to learn modification-specific features that are invariant to technical noise sources.

Class imbalance poses a particular challenge for modifications with low stoichiometry, where unmodified examples vastly outnumber modified ones. Strategies to address this include oversampling modified reads, weighted loss functions that penalize misclassification of the minority class more heavily, and focal loss which down-weights well-classified examples to focus learning on hard cases (Lin et al., 2017). For m^5C detection, where modification fractions are often below 50%, appropriate handling of class imbalance is essential to prevent models from defaulting to the unmodified prediction.

While these approaches improve model robustness, they address confidence estimation rather than the underlying decision boundary misalignment that generates false positives, a distinction that motivates the hard negative mining approach introduced in the following section.

1.5.6 Hard Negative Mining: A Principled Framework for Decision Boundary Refinement

Neural network classifiers implicitly construct decision boundaries in high-dimensional feature space. When training data inadequately represent the diversity of negative examples, these boundaries may be positioned incorrectly, leading to systematic false-positive predictions (Sung & Poggio, 1998). In RNA modification detection, an "unmodified" signal could arise from thousands of distinct sequence contexts, each with characteristic current profiles influenced by nucleotide composition, RNA secondary structure, and stochastic translocation dynamics. No training set can exhaustively sample this negative space.

Hard negative mining addresses this limitation through a key insight. The most informative negative examples for refining decision boundaries are those that the current model misclassifies. This iterative process is illustrated schematically in Figure 1.8. These "hard negatives" occupy regions of feature space near the decision boundary where discrimination is weakest. By identifying misclassified negatives and incorporating them into subsequent training iterations, the model progressively refines its boundaries. This concept traces to the work of Sung and Poggio (1998) on face detection. Their "bootstrapping" procedure established the paradigm of training an initial classifier, applying it to unlabeled data to identify false positives, adding these as negative examples, and retraining until performance stabilizes. Rowley et al. (1998) extended this to neural network-based face detection, demonstrating bootstrap training over 146 million candidate image windows. Shrivastava et al. (2016) formalized the approach for

modern deep learning with Online Hard Example Mining (OHEM), which ranks candidate regions by classification loss and backpropagates gradients only through the highest-loss examples. The theoretical justification derives from the geometry of classification boundaries. Hard negatives concentrate learning signal precisely where it is most needed (Robinson et al., 2021).

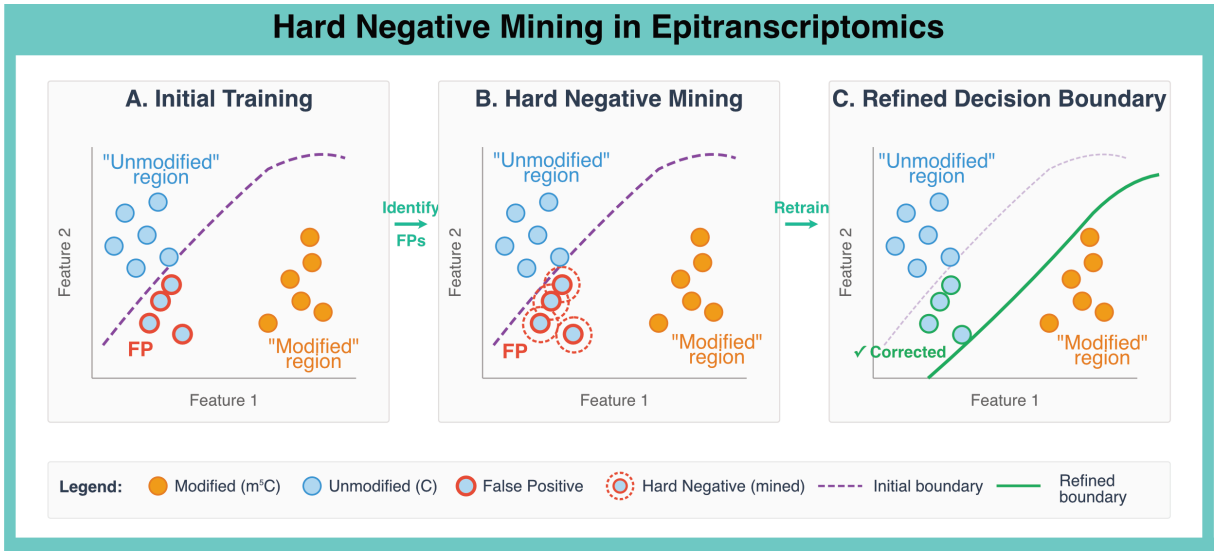


Figure 1.8. Hard negative mining for decision boundary refinement in epitranscriptomic classification. (A) Initial training produces a decision boundary from balanced modified (m^5C) and unmodified (C) data, but misclassifies certain unmodified positions as modified (false positives, red circles). (B) False positives are identified as hard negatives and added back to the training set as confirmed negative examples using IVT-derived ground truth. (C) Retraining with the augmented dataset yields a refined decision boundary that eliminates false positives while preserving true positive detection.

In medical imaging, hard negative mining has shown to be particularly valuable where false positives carry significant consequences and negative case diversity is substantial. Hard negative mining protocols have reduced false-positive rates by over 90% in lymphatic invasion detection while improving sensitivity (Lee et al., 2024). Pulmonary nodule detection presents challenges analogous to RNA modification detection, rare true positives, diverse false positives, and subtle distinguishing signals, where iterative hard negative mining has enabled clinical-grade performance (Tang et al., 2018). Notably, Xia et al. (2022) demonstrated that naive hard negative mining can backfire when hard negatives are actually false negatives, underscoring the importance of ground truth quality.

Despite the demonstrated power of hard negative mining across computer vision, medical imaging, and representation learning, this paradigm has not been systematically applied to RNA modification detection by nanopore sequencing. The field has instead focused predominantly on architectural innovation such as deeper networks, attention mechanisms, transformer layers, while training strategies that address decision boundary refinement have received comparatively

little attention. Yet as the preceding sections illustrate, the false-positive problem in m^{5}C detection might stem not from insufficient model capacity but from inadequate sampling of the heterogeneous negative space during training. Architectural complexity cannot compensate for training distributions that fail to represent the diversity of unmodified sequence contexts a model will encounter at deployment.

Hard negative mining offers a principled solution to this training distribution problem. By iteratively identifying false positives and incorporating them as negative examples, the approach directly addresses decision boundary misalignment rather than attempting to overwhelm it with model capacity. The challenge for nanopore applications lies in obtaining ground truth. In computer vision, false positives can often be verified by human inspection, a predicted face is either a face or not. In RNA modification detection, the modification status of a given position may be genuinely unknown, complicating the identification of true false positives versus unvalidated true positives. This challenge can be addressed using IVT RNA, which provides enzymatically guaranteed unmodified reference material and supplies the ground truth necessary for systematic hard negative mining.

The conceptual framework established here, treating false positives as diagnostic signals, using IVT-derived ground truth to identify hard negatives, and iteratively retraining to refine decision boundaries, provides the theoretical foundation for the calibration-based validation approach developed in this thesis.

1.5.7 The Validation Gap: From Discovery to Confirmation

Despite advances in *de novo* discovery tools, a critical gap remains: the absence of tools specifically designed for validation of the candidate sites. Existing deep learning approaches have been optimized for transcriptome-wide scanning, nominating candidate modification sites at scale. However, a fundamental arithmetic problem underlies this limitation. Mammalian transcriptomes contain millions of potential modification sites. Even a classifier achieving 99.9% specificity would generate thousands of false-positive calls at transcriptome scale, overwhelming true positives for low-abundance modifications like m^{5}C. Comparative benchmarking confirms that existing tools struggle to balance precision and recall for this modification (Zou et al., 2025; Luo et al., 2025). Indeed, a 2026 comprehensive review cataloguing all major deep learning approaches for nanopore RNA modification detection, including multi-modification profilers, specialized learning frameworks, and ensemble strategies, classified every tool exclusively within the discovery paradigm, repeatedly concluding that results require validation with orthogonal

biological assays (Ling et al., 2026). No reviewed method addressed validation as a distinct computational objective.

This reality motivates a complementary map-and-validate paradigm. Global scanners nominate candidates across the transcriptome, and calibration-based tools provide high-confidence validation for a selected subset where site-specific evidence is required. Rather than attempting to build universal classifiers that perform adequately across all contexts, the required precision can be achieved by narrowing the operational scope to specific, biochemically nominated sites. This thesis addresses the validation gap by developing such a calibration-based approach for site-specific m^5C confirmation using Nanopore DRS and deep learning.

2. Motivation and Objectives

As outlined in the preceding chapters, the growing repertoire of transcriptome-wide profiling methods has substantially accelerated the nomination of candidate RNA modification sites. Yet each of these approaches infers modification status through indirect biochemical readouts, chemical conversion, immunoprecipitation enrichment, or differential cleavage, and therefore carries method-specific biases that can generate false-positive signals, particularly for modifications with subtle signatures such as Nm and m^5C. This gap between candidate nomination and high-confidence confirmation has motivated a map-and-validate paradigm in the field, in which orthogonal, site-specific validation complements global profiling.

This project was originally conceived around dengue virus (DENV) genomic RNA as the primary model system for studying flaviviral RNA modifications; however, the onset of the SARS-CoV-2 pandemic introduced an urgent additional target. When the project began, Illumina-coupled approaches such as RiboMethSeq were the dominant framework for site-specific Nm analysis. Preliminary RiboMethSeq profiling of SARS-CoV-2 gRNA, performed in collaboration with the laboratories of Alessia Ruggieri (Heidelberg, Germany) and Yuri Motorin (Nancy, France), had nominated putative Nm sites on viral transcripts. The initial aim of this thesis was to validate two of these candidates using splint-ligated synthetic constructs carrying defined modification states. This validation effort, however, exposed systematic false-positive signals inherent to the cleavage-based detection principle, arising from RNA secondary structure and ligation biases. The inability to distinguish genuine modifications from technical artifacts using indirect readouts prompted a fundamental shift toward nanopore DRS, which provides a direct physical measurement of modification-induced current perturbations.

This transition, however, introduced its own practical challenges. Nanopore DRS was developed as a long-read sequencing platform optimized for full-length, polyadenylated transcripts, yet the field lacked systematic protocols for sequencing short, capped, or chimers RNAs, precisely the input diversity required for site-specific modification studies. Generating synthetic constructs that carry a defined modification at a single position and are compatible with the nanopore library preparation workflow demanded the development of robust ligation strategies, which constituted the first experimental objective of this thesis.

Among the modifications detectable by nanopore DRS, m^5C represents the most demanding target due to its minimal effect on ionic current, making it an ideal proving ground for a

validation framework. During this period, a collaborative effort between the Helm and Gerber groups yielded ModiDeC, a multi-modification deep learning classifier originally designed to detect m^6A, inosine, m^1A, and pseudouridine from nanopore DRS data using site-specifically modified training constructs (Alagna et al., 2025). Building on ModiDeC, a central objective of this thesis was to retrain this classifier for m^5C detection on viral RNA and to develop a calibration strategy, termed ModiCal, that systematically eliminates the false-positive predictions that are characteristic of this modification across existing detection approaches. Using the well-characterized m^5C sites on *Saccharomyces cerevisiae* 25S rRNA as the initial biological model, ModiCal implements iterative retraining to refine decision boundaries toward site-specific, validation-grade precision. The generalizability of this approach was tested by applying it to the original viral target, DENV genomic RNA, which carries a low-stoichiometry m^5C deposited by NSUN6.

3. Results and Discussion

3.1 Ligation Strategies for Site-Specifically Modified RNA

The initial phase of this thesis (2022–2024) explored different splint-mediated ligation strategies for generating site-specifically modified RNAs suitable for RNA modification mapping and validation. These experiments encompassed three principal activities: (A) splint-mediated ligation of long and short constructs bearing defined 2'-O-methylation sites, for the validation of RiboMethSeq analysis of SARS-CoV-2 candidate sequences; (B) development of a chemical capping protocol with potential application in modification detection workflows; and (C) functional characterization of modification effects on translation using *in vitro* translation assays as an application. While none of these activities produced the originally anticipated biological findings, they yielded critical technical insights, most notably, the discovery of systematic false-positive signals in RiboMethSeq and the demonstration that short splint ligation constructs are fully compatible with nanopore DRS, that directly motivated the development of the ModiCal calibration framework described in Section 3.2.

3.1.1 Validation of SARS-CoV-2 2'-O-Methylation Candidates by RiboMethSeq

3.1.1.1 Context and Rationale

Preliminary RiboMethSeq profiling of SARS-CoV-2 viral genomic RNA isolated from infected cells by the laboratory of Alessia Ruggieri (Heidelberg University) and analyzed in collaboration with the laboratory of Yuri Motorin (Université de Lorraine, Nancy, France), had identified multiple candidate Nm positions across the viral genome. From this dataset, 34 high-confidence sites were selected out of 80 common positions found in two independent SARS-CoV-2 samples (ScoreMEAN ≥ 0.92, minimum ScoreA2 ≥ 0.5) (Figure 3.1A). Subsequent re-analysis using a Random Forest classification model (Pichot et al., 2022) refined the candidate list to 9 trustworthy putative methylated sites and 35 additional candidates requiring further validation.

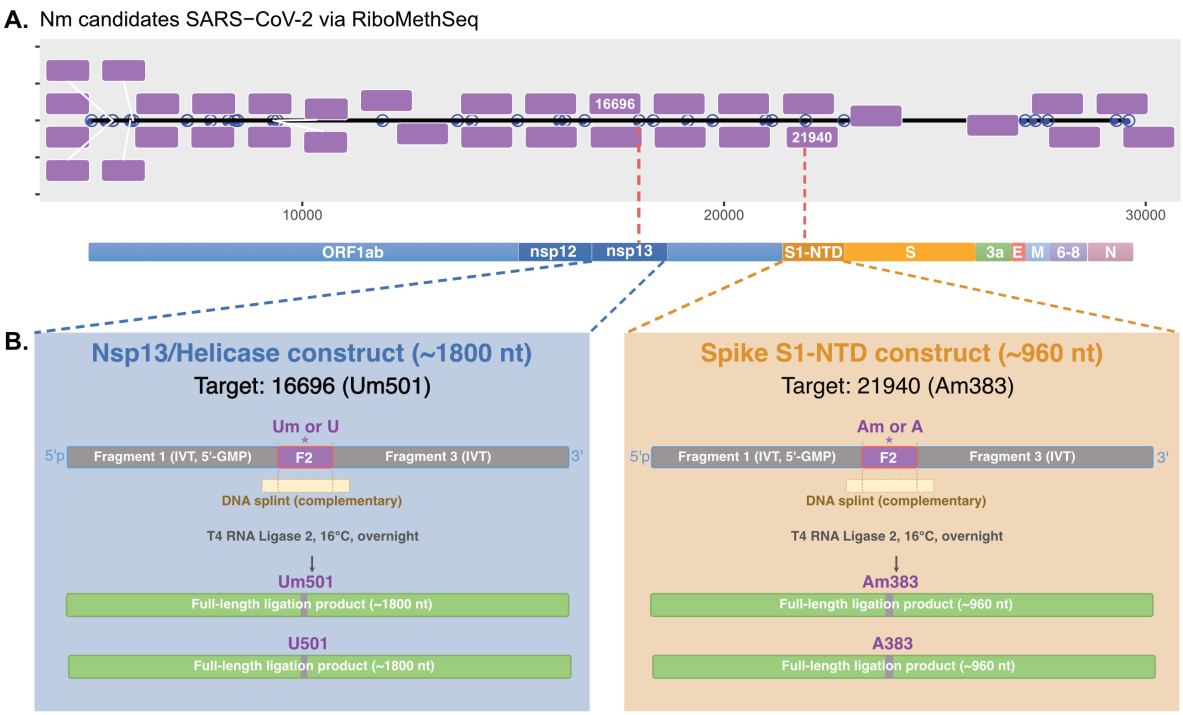


Figure 3.1 RiboMethSeq-derived Nm candidates on SARS-CoV-2 and splint ligation validation strategy. (A) Selected list of 34 high-confidence Nm candidates (purple) out of 80 common positions found in two SARS-CoV-2 samples (LimMEAN ≥ 0.92, min A2 ≥ 0.5). Positions 16696 (Um501, Nsp13) and 21940 (Am383, Spike S1-NTD) were selected for experimental validation. **(B)** Three-fragment splint ligation strategy: a central synthetic oligoribonucleotide (F2) bearing either the 2'-O-methylated or unmodified nucleotide was ligated between two IVT-derived flanking fragments using T4 RNA Ligase 2, yielding modified and unmodified constructs for RiboMethSeq analysis.

Two candidate sites were selected for experimental validation based on their high confidence scores, their location within functionally relevant coding regions, and practical constraints of the splint ligation strategy, which requires the modification site to fall near the midpoint of the target region to satisfy size-symmetry requirements (see Section 3.1.1.2):

1. **2'-O-methyluridine at genome position 16696**, corresponding to relative position 501 (Um501) within the helicase Nsp13 coding region.
2. **2'-O-methyladenosine at genome position 21940**, corresponding to relative position 383 (Am383) within the Spike protein S1 N-terminal domain (S1-NTD) coding region.

The validation strategy was based on comparing the RiboMethSeq profiles of synthetic constructs bearing defined modification states at the candidate positions. If the candidate sites were genuinely methylated in the viral RNA, then constructs carrying the corresponding Nm should produce a distinct protection signal relative to unmodified controls. Conversely, if the RiboMethSeq signals were technical artifacts, modified and unmodified constructs would yield indistinguishable profiles.

3.1.1.2 Splint-Mediated Ligation of SARS-CoV-2 Constructs

To generate RNA molecules bearing defined Nm at precise positions, a three-way one-pot splint-mediated ligation strategy was employed, based on the protocol established in this laboratory (Hertler, Slama et al., 2022; Moore & Sharp, 1992; Kershaw & O'Keefe, 2012). In this approach, a full-length RNA is assembled from three fragments: two flanking fragments (Fragment 1 and Fragment 3) produced by IVT, and a central chemically synthesized oligoribonucleotide (Fragment 2) containing the modification of interest at a defined position. The three-fragment design requires that Fragment 1 and Fragment 3 be of approximately equal length, such that the full-length ligation product migrates at roughly twice the size of the individual fragments on a preparative gel, enabling clear separation of ligated from unligated material. A complementary DNA oligonucleotide (the splint) bridges the junctions between adjacent fragments, enabling enzymatic ligation by T4 RNA Ligase 2, which preferentially joins RNA ends within an RNA/DNA duplex context (Nandakumar et al., 2004).

Two SARS-CoV-2 target regions were assembled:

1. **Spike S1-NTD construct** (~960 nt): targeting Am383. Fragment 1 (F1) was generated by IVT with 5'-GMP initiation to provide the requisite 5'-monophosphate terminus for ligation. Fragment 2 (F2, 19 nt synthetic oligoribonucleotide) was either unmodified or bearing Am at position 383. Fragment 3 (F3) was generated by IVT.
2. **Nsp13/Helicase construct** (~1800 nt): targeting Um501, with an analogous three-fragment strategy.

Plasmid templates for IVT of the individual fragments were synthesized by GenScript in pUC57 backbones. Construct design, fragment boundaries, and oligonucleotide sequences for the Nsp13 construct are published in Hertler, Slama et al. (2022) (Figure 3.2A).

Equimolar amounts of Fragment 1, the phosphorylated oligoribonucleotide (F2), and Fragment 3 were combined with the DNA splint in T4 RNA Ligase Buffer, and ligated overnight. Following DNase I digestion to remove the DNA splint, ligation products were separated by real-time gel elution (RTGE) on preparative 1% agarose gels containing 1× SYBR Gold, adapted from the E-Gel CloneWell concept (Hertler, Slama et al., 2022). The full-length ligation band was collected from the elution well, and RNA was recovered by ammonium acetate/ethanol precipitation and filtered through 0.2 µm NanoSep centrifugal filters to remove residual agarose (Figure 3.2B).

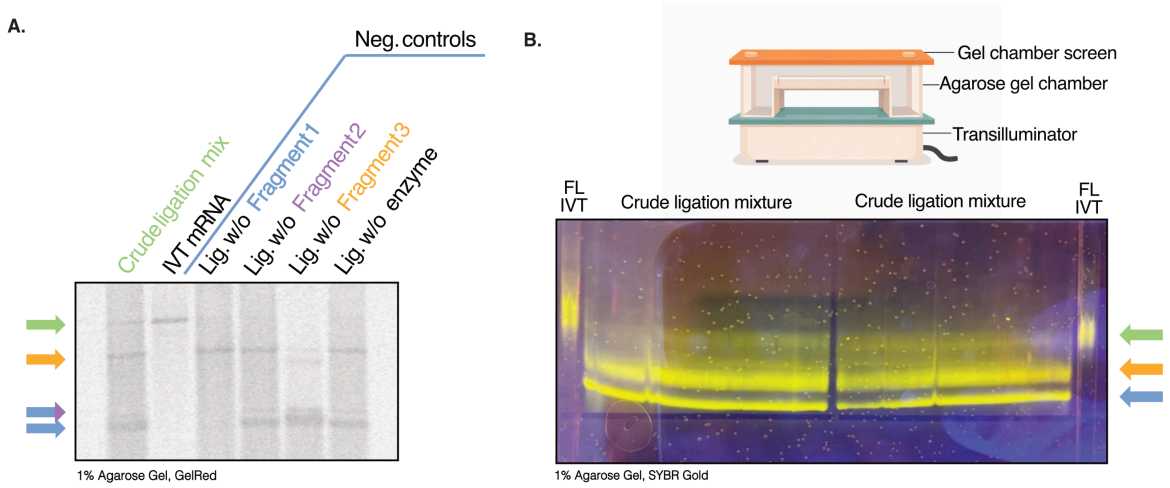


Figure 3.2. Splint-mediated ligation and purification of the Nsp13/helicase construct (~1800 nt). (A) Analytical 1% agarose gel, GelRed staining. Lane 1: crude ligation mix; Lane 2: full-length IVT control. Lanes 3–6: negative controls omitting Fragment 1, Fragment 2, Fragment 3, or T4 RNA Ligase 2, respectively. Adapted from Hertler, Slama et al. (2022). Experiments performed by the author. **(B)** Preparative 1% agarose gel containing 1× SYBR Gold, visualized on a Dark Reader Blue Light Transilluminator for real-time gel elution (RTGE). FL IVT: full-length IVT size reference. Green arrow: full-length ligation product; orange arrow: Fragment 3; blue arrow: Fragment 1; purple arrow (A only): Fragment 2 (19 nt). Gel chamber and blue light transilluminator illustrations in panel were modified from Hertler, Slama et al. (2022) using ChatGPT.

The following sample set was generated for RiboMethSeq analysis:

Table 3.1 Sample overview for RiboMethSeq validation

Sample ID	Construct	Description	Modification
MH380	Spike S1-NTD	Full-length IVT	None (IVT control)
MH381	Spike S1-NTD	Unmodified ligation	None (ligation control)
MH382	Spike S1-NTD	Modified ligation	Am383
MH383	Nsp13	Full-length IVT	None (IVT control)
MH384	Nsp13	Modified ligation	Um501

For the Nsp13 construct, an unmodified ligation control analogous to MH381 was not available as the gel elution step for this sample failed to yield sufficient material for RiboMethSeq analysis. The full-length IVT control (MH383), which carries no modifications, was considered a sufficient negative control for the purposes of assessing false-positive signals.

3.1.1.3 RiboMethSeq Analysis and the Discovery of Systematic False Positives

Before reanalysing the candidate sites with their known modification status, it is necessary to revisit the reliability of threshold-based RiboMethSeq calling. Pichot et al. (2020) demonstrated that the original ScoreMEAN calculation with a ±6 nt window yielded a false discovery rate (FDR) of 27% in RiboMethSeq datasets, which could be reduced to 4% by optimizing the

calculation interval to ±2 nt. Even with this optimization, the absolute number of false positives remains substantial at the transcriptome-wide level. The ScoreC2 metric is known to behave abnormally in regions of highly irregular fragmentation profiles, producing negative values that are theoretically impossible for a ratio-based score (Marchand et al., 2016). Approximately 10% of known methylated positions in yeast rRNA showed ScoreC2 differences between modified and unmodified RNA as low as 0.1, making reliable detection inherently difficult at these sites (Pichot et al., 2020). Furthermore, pseudouridine sites can generate fragment end gaps that mimic Nm signals, constituting an additional source of false positives (Pichot et al., 2020).

The five constructs listed in Table 3.1 were subjected to RiboMethSeq analysis in the laboratory of Yuri Motorin following the established Illumina-based protocol (Marchand et al., 2016). ScoreC2 values were computed at each position using a ±2 nt calculation interval.

The analysis yielded unexpected results. The global RiboMethSeq fragment end count profiles across the entire length of each construct were inspected prior to position-specific analysis (Figure 3.3). Both the Spike S1-NTD (~960 nt) and Nsp13 (~1800 nt) constructs displayed highly irregular fragmentation patterns, with coverage varying over several orders of magnitude. Crucially, these irregular profiles were already present in the full-length IVT controls, which carry no modifications at any position. This is expected as apart from the single candidate position under investigation, the entire sequence is unmodified in all samples, and therefore the fragmentation pattern should, and does, reflect intrinsic sequence and structure properties rather than modification status. Importantly, the comparison between IVT and ligation samples revealed that the splint ligation procedure itself does not introduce detectable distortions in the overall RiboMethSeq fragmentation profile. The global coverage patterns of ligation products closely matched those of the corresponding full-length IVT controls, indicating that the ligation junctions do not cause widespread artifacts. Any differences between samples are therefore confined to the local sequence context around candidate positions, as examined below.

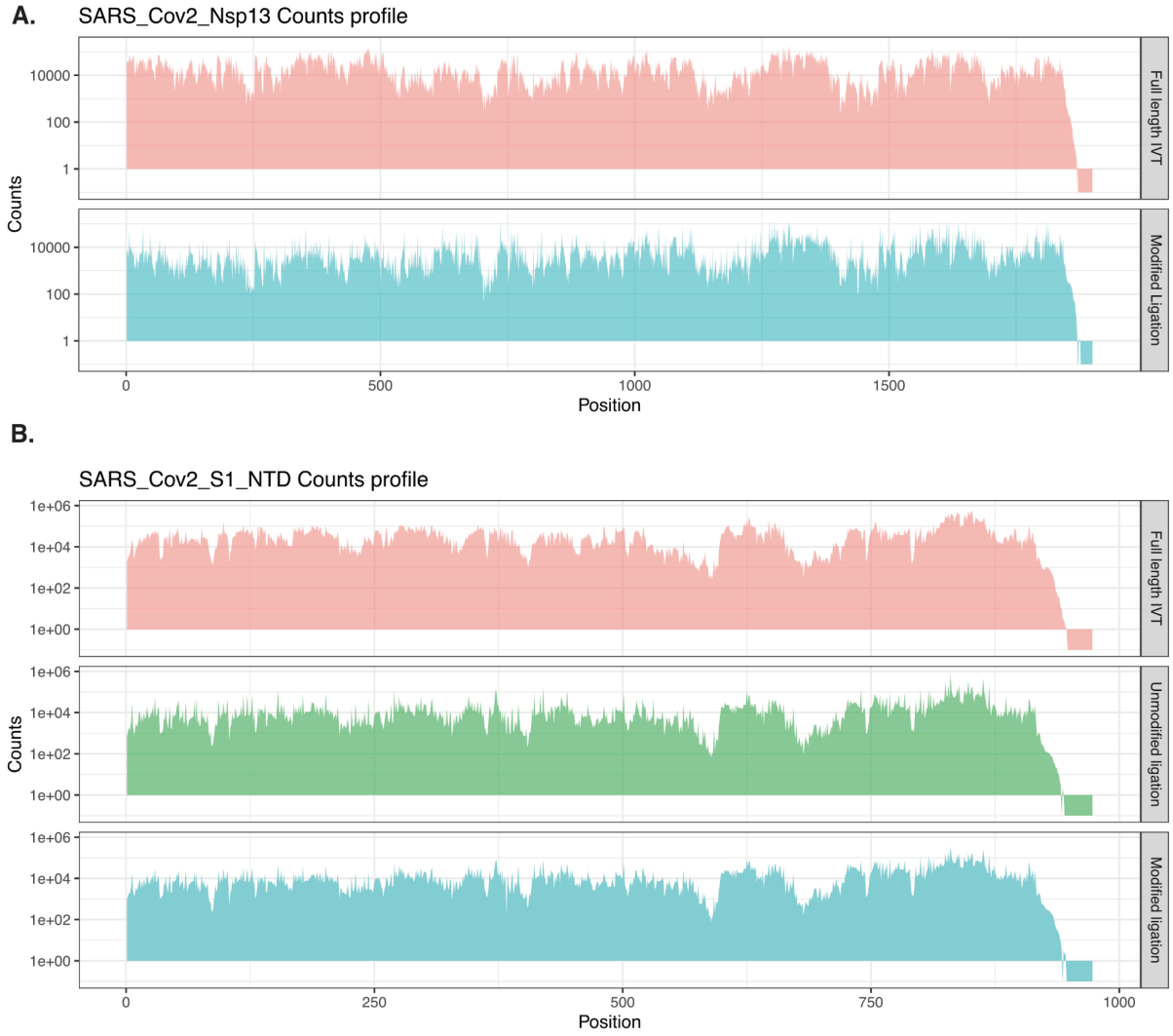


Figure 3.3 RiboMethSeq fragment end count profiles across the full length of the ligation constructs. (A) Nsp13/helicase construct (~1800 nt). Top: full-length IVT, unmodified; bottom: Um501-modified ligation. Y-axis: fragment end counts (log_{10} scale). **(B)** Spike S1-NTD construct (~960 nt). Top: full-length IVT, unmodified; middle: unmodified ligation; bottom: Am383-modified ligation.

While the modified ligation products showed the anticipated protection signals at the engineered modification sites, the unmodified controls also exhibited substantial protection at the same positions, indicating false-positive protection signals, since these control samples contain no 2'-O-methylation by design.

Table 3.2 RiboMethSeq ScoreC2 values for Nsp13 constructs

Sample	Description	ScoreC2 (pos. 501)	ScoreC2 (pos. 652)	Expected
MH383	Full-length IVT (unmodified)	0.828	0.779	~0
MH384	Modified ligation (Um501)	0.976	0.951	~1 (pos. 501), ~0 (pos. 652)

Table 3.3 RiboMethSeq ScoreC2 values for Spike S1-NTD constructs at position 383

Sample	Description	ScoreC2 (pos. 383)	Expected
MH380	Full-length IVT (unmodified)	0.712	~0
MH381	Unmodified ligation (no Nm)	0.888	~0
MH382	Modified ligation (Am383)	0.982	~1

For the Nsp13 construct at position 501, the unmodified IVT control, which by definition carries no 2'-O-methylation, already displayed a ScoreC2 of 0.828 (Figure 3.4A). This value substantially exceeds the standard ScoreC2 > 0.6 calling threshold, meaning that the unmodified IVT itself would be classified as "modified" by the method's own criteria, a result that is, by definition, a false positive. The Um501-modified ligation product reached 0.976, but the high baseline in the unmodified control precludes confident discrimination between genuine modification and artifact.

An additional observation reinforced the artifact interpretation. Position A652 on the Nsp13 construct, at which no modification had been introduced, was automatically detected as a high-confidence Nm candidate, exhibiting ScoreC2 values of 0.779 (IVT control) and 0.951 (modified ligation) (Figure 3.4B; Table 3.2). This constituted a clear false positive, as its high protection in the modification-free IVT control excluded genuine 2'-O-methylation as an explanation. The fact that an entirely unintended position produced ScoreC2 values exceeding the calling threshold in all samples known to be unmodified on that site, provides direct evidence that RiboMethSeq protection signals can arise independently of Nm.

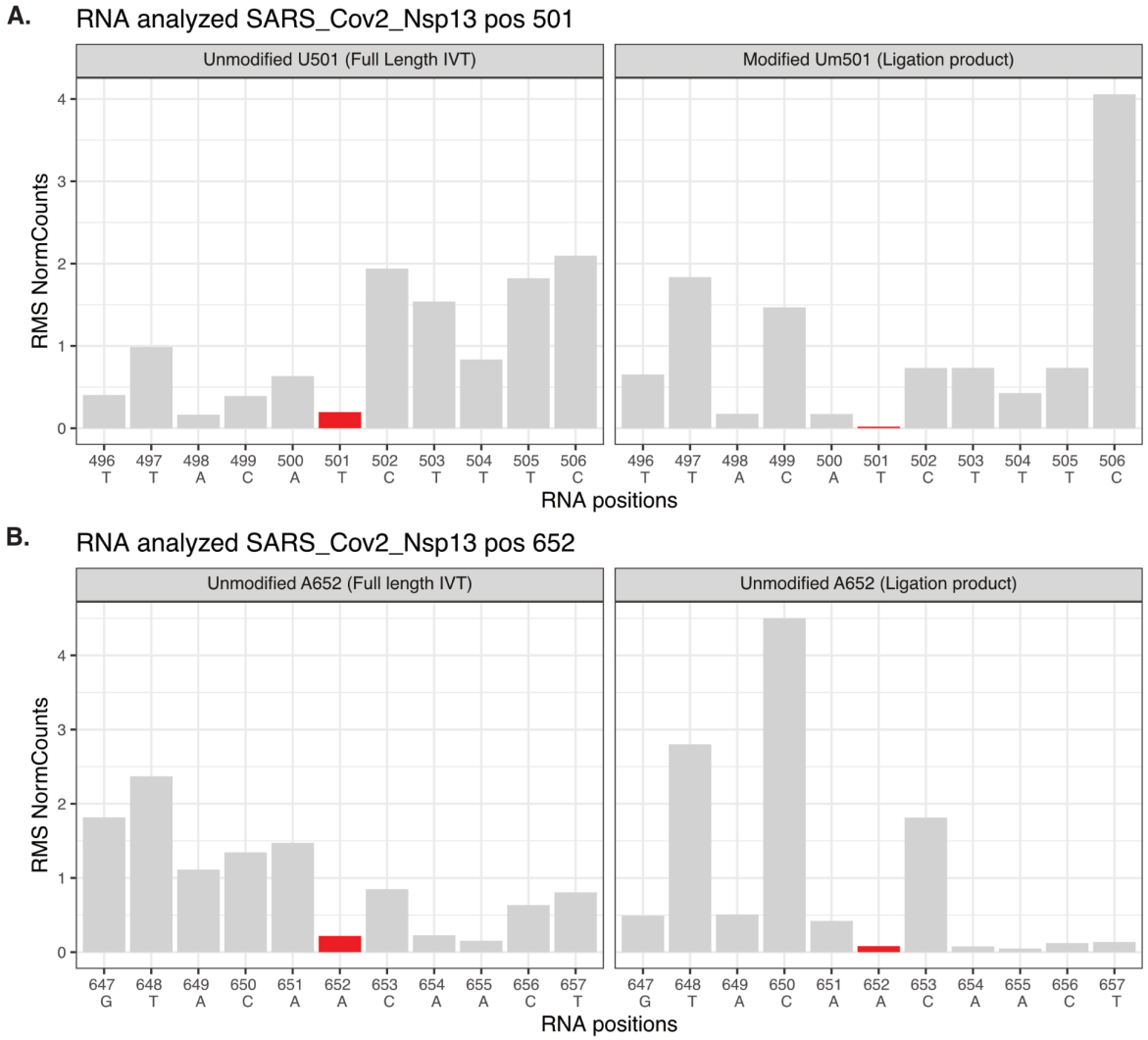


Figure 3.4. RiboMethSeq normalized count profiles for Nsp13/helicase constructs. (A) Position 501 (Um501 target site). Left: full-length IVT, unmodified, ScoreC2 = 0.828; right: Um501-modified ligation, ScoreC2 = 0.976. **(B)** Position 652 (no modification introduced at this site). Left: IVT, ScoreC2 = 0.779; right: ligation product, ScoreC2 = 0.951. The red bar indicates the candidate position.

The same pattern was confirmed for the Spike S1-NTD construct at position 383. The full-length IVT control exhibited a ScoreC2 of 0.712, again above the 0.6 threshold, while the unmodified ligation product showed an even higher value of 0.888 and the genuinely modified construct reached 0.982 (Figure 3.5; Table 3.3). The normalized count profiles at position 383 visually illustrate this phenomenon. All three samples show reduced fragment end counts at the candidate position, with the degree of protection escalating from IVT to unmodified ligation to modified ligation. Notably, the stepwise increase from IVT (0.712) to unmodified ligation (0.888) suggests that the ligation process itself introduces additional local structural perturbations that affect cleavage patterns, independent of any modification.

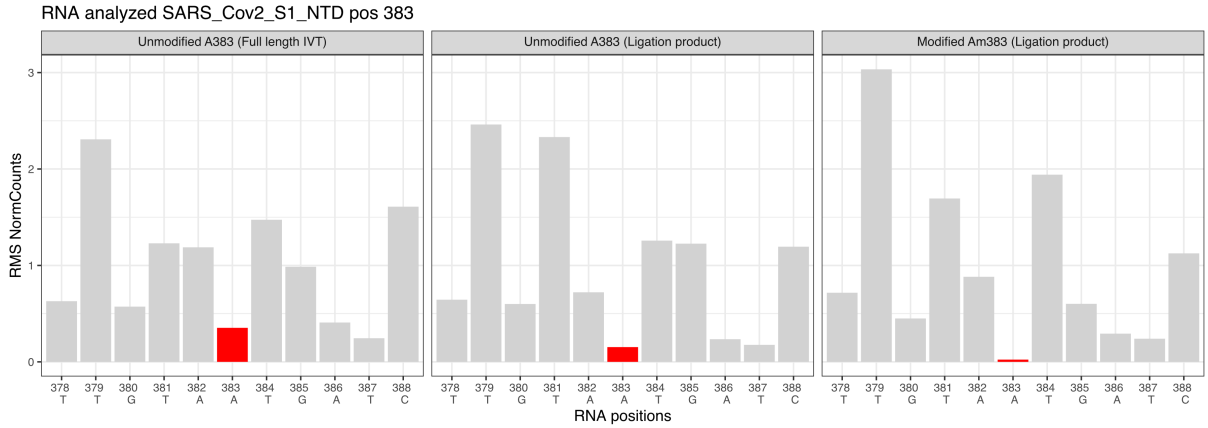


Figure 3.5. RiboMethSeq normalized count profiles at position 383 of the Spike S1-NTD construct. Left: full-length IVT, unmodified, ScoreC2 = 0.712. Center: unmodified ligation, ScoreC2 = 0.888. Right: Am383-modified ligation, ScoreC2 = 0.982. The red bar indicates the candidate position.

The observation of substantial protection signals in samples known to be unmodified indicates the presence of systematic technical artifacts in RiboMethSeq. Several non-mutually exclusive mechanisms can account for this phenomenon:

RNA secondary structure. Local secondary structures can protect phosphodiester bonds from alkaline hydrolysis independently of modification status. If a nucleotide is constrained within a stable stem-loop or base-paired region, reduced solvent accessibility diminishes cleavage efficiency and produces a protection signal indistinguishable from genuine Nm (Helm & Motorin, 2017; Motorin & Marchand, 2018). Highly structured RNA regions require approximately 20-fold more sequencing reads than theoretically predicted for reliable coverage due to irregular fragmentation patterns (Motorin & Marchand, 2018). Both the Spike S1-NTD and Nsp13 sequences are predicted to form stable secondary structures, consistent with this mechanism.

Sequence-dependent cleavage bias. The efficiency of alkaline hydrolysis varies with local sequence context. Certain dinucleotide and trinucleotide combinations exhibit intrinsically reduced cleavage rates, creating position-specific biases that manifest as elevated ScoreC2 values in the absence of any modification. This effect has been documented as a significant source of error in RiboMethSeq datasets, particularly when analyzing non-ribosomal RNAs where the modification landscape is not independently established (Helm & Motorin, 2017; Pichot et al., 2020).

Ligation junction effects. The notably higher ScoreC2 observed in the unmodified ligation product (MH381, 0.888) compared to the full-length IVT (MH380, 0.712) at the same position suggests that the splint ligation process itself introduces local structural perturbations that affect

subsequent cleavage patterns. This could reflect altered folding near the ligation junction or residual effects of the DNA splint annealing step on RNA conformation.

Density of false-positive signals. The false-positive problem observed here is not isolated. In the original preliminary RiboMethSeq profiling of SARS-CoV-2 RNA from infected cells, 80 positions showed protection signals common to two independent samples, of which only 34 passed the stringent filtering criteria (ScoreMEAN ≥ 0.92, ScoreA2 ≥ 0.5). Even among these, a refined analysis using a Random Forest machine learning classifier, which was specifically developed to reduce false-positive rates compared to threshold-based approaches (Pichot et al., 2022), retained only 9 positions as "trustworthy" candidates. The observation that the vast majority of initial candidates were likely artifacts underscores the challenge of Nm site identification in structured viral RNAs.

These results had two principal consequences:

First, the putative 2'-O-methylation sites on SARS-CoV-2 Spike (Am383) and Nsp13 (Um501) RNA that had been nominated by RiboMethSeq were validated as false positives.

Second, the false-positive problem revealed by these experiments illustrates a challenge that extends beyond RiboMethSeq to any modification detection method relying on indirect biochemical readouts. Cleavage-based, antibody-based, and chemical conversion-based methods are all susceptible to context-dependent artifacts that can mimic genuine modification signals (Helm & Motorin, 2017; Motorin & Helm, 2024). The experience gained here, that false positives are not random noise but arise from systematic, context-dependent features of the detection method, informed the broader analytical approach taken in subsequent chapters of this thesis.

3.1.1.4 Transition to Short Splint Ligation for Nanopore Sequencing

The mapping validation strategy was therefore shifted from cleavage-based modification detection toward nanopore DRS, which reports modifications through real-time changes in ionic current rather than through indirect biochemical readouts. On the other hand, the splint ligation of long IVT-derived constructs proved impractical. Yields after gel purification were extremely low, and for the Nsp13 construct the unmodified ligation control could not be recovered at all, as the gel elution step failed to produce sufficient material for downstream analysis (Table 3.2). Together, these observations prompted a systematic evaluation of whether short ligation constructs could serve as an efficient and nanopore-compatible source of ground-truth training data for modification detection.

To this end, a systematic comparison of ligation efficiency was conducted using the IRESeGFP5' test construct, a 1371-nucleotide model mRNA consisting of a 5' UTR with an internal ribosome entry site (IRES) followed by the enhanced green fluorescent protein (eGFP) coding sequence, which had been established as a model system for previous splint ligation optimization (Hertler, Slama et al., 2022). For the long construct, the full-length mRNA was assembled from three fragments. Two IVT-derived flanking fragments (F1: ~630 nt, F3: ~720 nt) and a central 19-nucleotide synthetic oligoribonucleotide (F2) carrying the desired 2'-O-methylation at a defined position near the AUG start codon (Figure 3.6A). For the short construct, the same F2 oligo was flanked by truncated 40-nucleotide fragments, yielding a 99-nucleotide ligation product (Figure 3.6A). Denaturing PAGE analysis confirmed that the short ligation products could be clearly resolved from their constituent fragments, with the full-length 99 nt product, partial ligation intermediates (59 nt), and individual fragments (40 nt, 19 nt) migrating as distinct bands (Figure 3.6B).

Yield quantification demonstrated that short constructs were consistently at least three-fold more efficient than long constructs across all modification types tested, including unmodified, Gm (position 12), Um (position 13), and Gm (position 16) within the F2 oligo (Figure 3.6C). Yields were calculated as the ratio of purified product mass to total input RNA mass. Short ligation products were recovered at yields of 3–6 µg per reaction, whereas long construct recovery was substantially lower, in some cases as low as 1% of input material. This improvement is consistent with the reduced conformational complexity of short oligonucleotides, which facilitates annealing to the DNA splint without competing intramolecular secondary structures (Kershaw & O'Keefe, 2012).

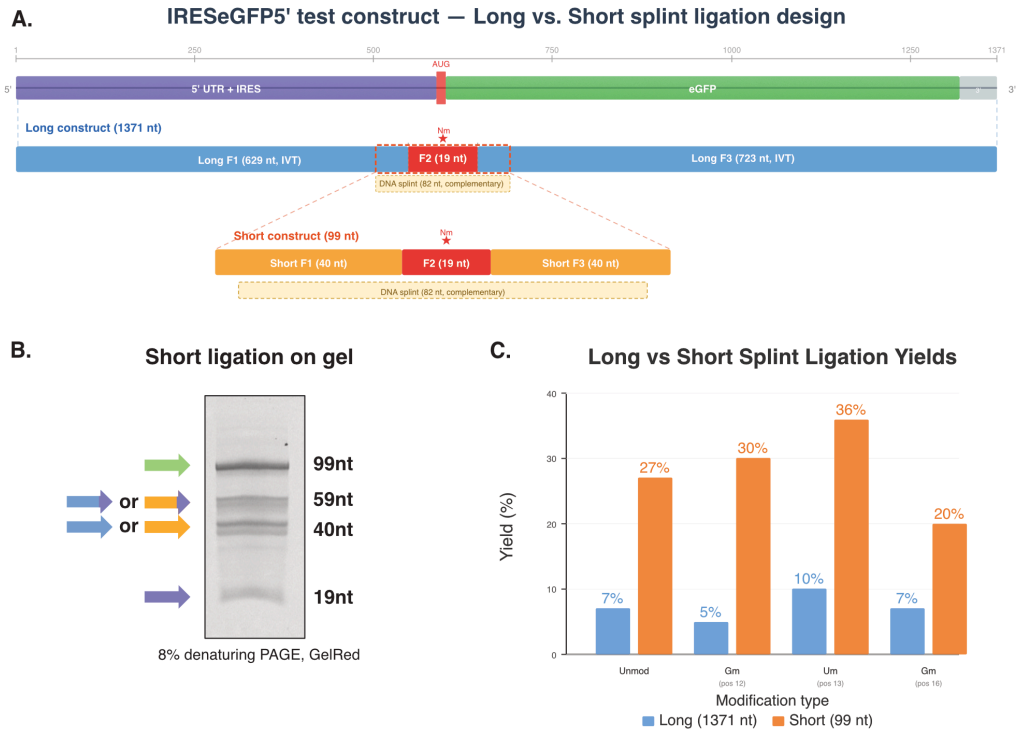


Figure 3.6 Long versus short splint ligation of the IRESeGFP5' test construct. (A) Schematic of the IRESeGFP5' mRNA (1371 nt) showing the 5' UTR with IRES (purple), AUG start codon (red), and eGFP coding sequence (green). The long construct is split into F1 (629 nt, IVT), F2 (19 nt, synthetic oligo with Nm), and F3 (723 nt, IVT); the short construct uses truncated flanking fragments (F1, F3: 40 nt each) yielding a 99 nt product. (B) Denaturing PAGE (8%, GelRed) of the short ligation reaction. Green arrow: full-length product (99 nt); blue/orange arrows: partial ligation intermediates (59 nt) and individual fragments (40 nt); blue arrow: F2 oligo (19 nt). (C) Yield comparison for unmodified, Gm (position 12), Um (position 13), and Gm (position 16) constructs. Short ligation consistently achieves 3–6× higher yields than long ligation.

Critically, the short ligation products were found to be fully compatible with nanopore DRS. After 3'-polyadenylation and adapter ligation, both long and short constructs were sequenced on Oxford Nanopore MinION flow cells. Comparison of the resulting alignment profiles revealed that short constructs provided uniform read coverage across the construct length, comparable to the long constructs, and that the modification-induced basecalling perturbation at the 2'-O-methylated position was clearly detectable in both formats (Figure 3.7A-B). This confirmed that the reduced construct length did not compromise modification detection sensitivity.

One technical observation from the nanopore sequencing data concerned the 5' end of the reads. DRS consistently failed to capture the terminal ~11 nucleotides at the 5' end of the construct (Figure 3.7B, red double-ended arrow). To determine whether this loss was an artifact of the ligation product or a general property of the sequencing technology, the same short construct was also sequenced using nanopore cDNA sequencing. The cDNA reads covered the full length of the construct including the 5' terminus, demonstrating that the truncation is intrinsic to the DRS process rather than a feature of the splint-ligated RNA (Figure 3.7B-C). This observation is consistent with the known limitation of nanopore DRS, in which the motor protein is unable to

fully read the terminal 10–15 nucleotides at the 5' end of the RNA strand as it transits the pore (Ibrahim et al., 2021). For the 99-nucleotide short constructs used in this work, this 5' loss is proportionally more significant (~11%) than for longer transcripts and was considered in the design of subsequent experiments (Section 3.2).

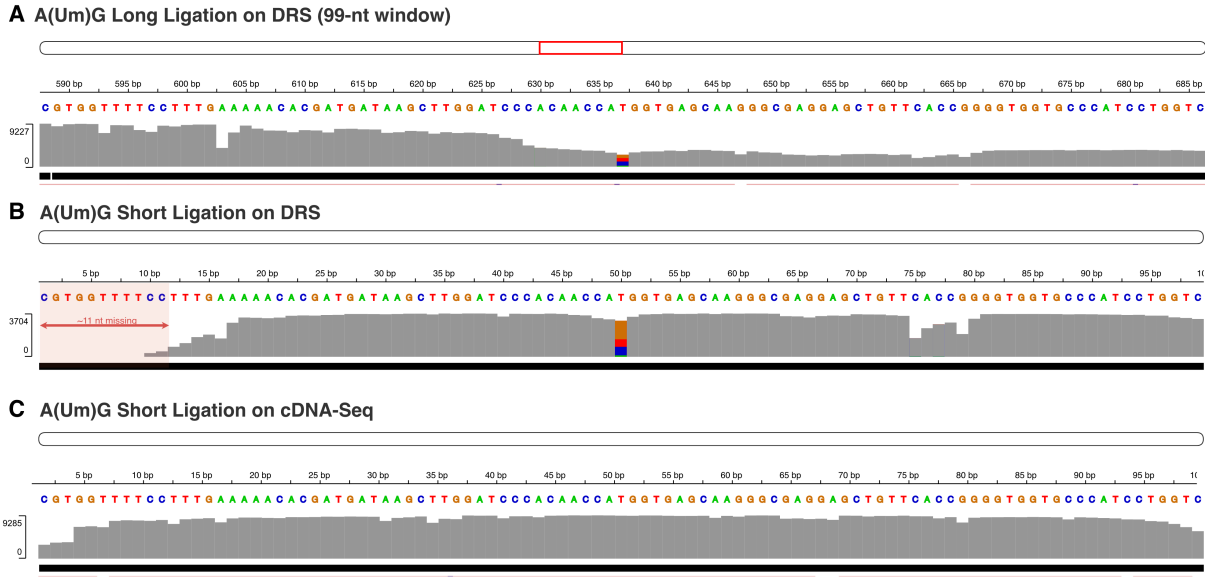


Figure 3.7. Nanopore sequencing alignment of the Um-modified IRESeGFP5' constructs. **(A)** DRS reads from the long construct (1371 nt), zoomed to the corresponding 99 nt region. Colored mismatches at the Um site (position 637) indicate modification-induced basecalling errors. **(B)** DRS reads from the short construct. The 5'-terminal ~11 nucleotides lack coverage (red), reflecting the intrinsic 5' truncation of nanopore DRS. **(C)** cDNA sequencing of the same short construct, showing full-length coverage including the 5' end and no mismatch signal at the modification site. Aligned with Minimap2; visualized in IGV.

3.1.2 Chemical Capping and Application to Translational Studies

3.1.2.1 Rationale

The long-construct ligation strategy described in Section 3.1.1, posed significant challenges for both RiboMethSeq validation and potential application in translational or immunostimulation assays. The IRES-containing IRESeGFP5' construct required three-fragment assembly and a total length exceeding 1300 nt, both of which substantially reduced ligation efficiency and limited the yield of full-length product available for downstream applications. An alternative strategy was therefore pursued: assembling shorter constructs from fewer fragments to improve ligation yield, while introducing a post-ligation chemical capping step to furnish the 5'-m^{7}G cap structure required for eukaryotic translation initiation. This approach enabled the systematic investigation of whether the position of an internal 2'-O-methylation within a coding sequence affects translational efficiency, as has been reported in a position-dependent manner (Choi et al., 2018; Hoernes et al., 2016). Although the translational assays did not ultimately converge with the nanopore-based modification detection pipeline, the chemical capping protocol developed

in this context holds potential for future application in modification detection workflows requiring defined 5'-cap structures.

To this end, a library of IRESeGFP5' mRNA constructs was produced, each bearing a single Nm at a different position within the same 19-nucleotide central oligoribonucleotide. The library comprised an unmodified control and ten modified variants carrying Am, Um, or Gm at each of the three positions within the start codon (AUG), the first coding codon (GUG, Val), and the second coding codon (AGC, Ser), plus one additional variant at the position preceding the start codon (AmCC). These constructs share the identical IRESeGFP5' backbone sequence (1371 nt), ensuring that any differences in translational output can be attributed to the position of the modification rather than sequence variation. The IRES element drives cap-independent translation initiation, while the eGFP coding sequence provides a quantifiable readout via fluorescence.

3.1.2.2 *In Vitro* Translation Results

In vitro translation was performed using the Flexi RRL system (Section 5.2.2). Preliminary screening of the full library by in-gel fluorescence detection (Slama, 2021) identified three constructs with the most pronounced effects on translation: AmUG (near-complete inhibition), GmUG and GUGm (apparent enhancement of ~2-fold) (Figure 3.8A). However, the in-gel fluorescence signals were very low even for positive controls, precluding reliable quantification. The replication assay was therefore transitioned to a Tecan plate reader-based fluorescence measurement, which provided markedly improved sensitivity and dynamic range. A 1:3 dilution series established that the Tecan system could reliably differentiate translation levels across a meaningful range.

Based on the preliminary screening, the three constructs showing the most biologically interesting effects (AmUG, GmUG, GUGm) were selected for validation, together with the unmodified control for normalization. The plate reader measurements confirmed that Nm modification on the first position of the start codon (AmUG) nearly completely inhibited translation, reducing output to approximately 20% of the unmodified control ($p \leq 0.001$; $n = 3$ biological replicates, three technical replicates each; Figure 3.8B). However, the two constructs that had shown ~2-fold enhancement in the preliminary data yielded conflicting results. Both GmUG and GUGm showed significant repression rather than enhancement, reducing translational output to approximately 70% and 64% of the unmodified control, respectively ($p \leq 0.01$; Figure 3.8B). Given these contradictory findings and the inherently low translational

efficiency of the IRES-dependent system, which limited the dynamic range available for detecting subtle modification effects, the project was not pursued further.

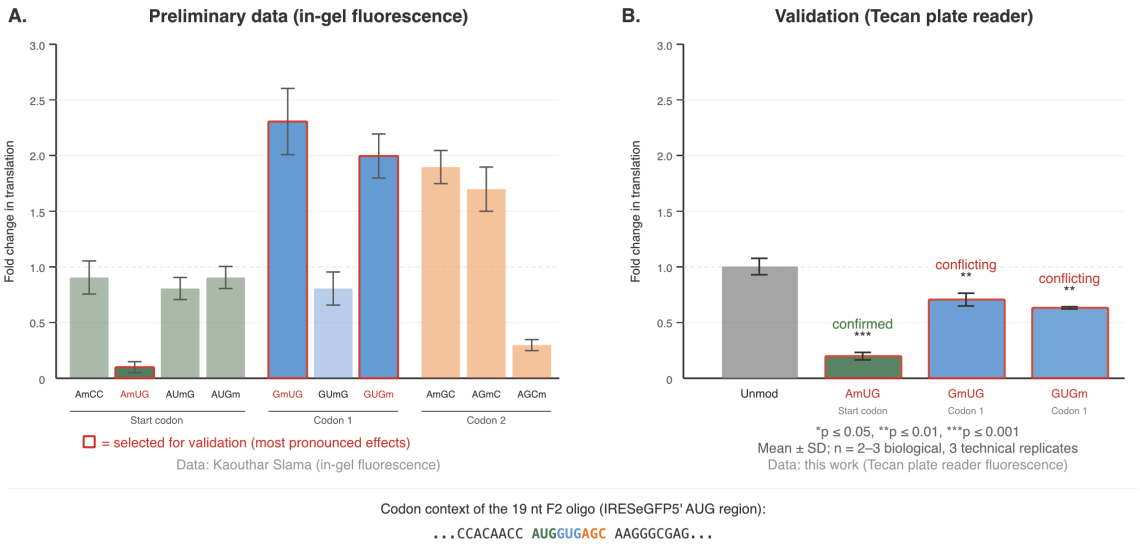


Figure 3.8. Effect of 2'-O-methylation position on IRES-dependent translation of IRESeGFP5' constructs. (A) Preliminary screening of the full position-scanning library by in-gel fluorescence (data: [redacted]). Ten Nm-modified constructs spanning the start codon (AmCC, AmUG, AUmG, AUGm), first coding codon (GmUG, GUmG, GUGm), and second codon (AmGC, AGmC, AGCm) are shown, normalized to the unmodified control (dashed line at 1.0). Red-outlined bars indicate the three constructs selected for replication based on their most pronounced effects. **(B)** Replication of selected constructs by Tecan plate reader fluorescence measurement (this work). AmUG (start codon position 1) nearly completely inhibits translation (confirmed); GmUG and GUGm (codon 1 positions 1 and 3) show repression rather than the enhancement observed in preliminary data (conflicting). Data represent mean ± SD; n = 2–3 biological replicates, three technical replicates each. **p ≤ 0.01, ***p ≤ 0.001 versus unmodified control. Panel A plot redrawn from Slama (2021).

3.1.2.3 Development of Chemical Capping

The low ligation efficiency of long IRES-containing constructs (as low as 1% recovery after gel purification) created a practical bottleneck for translation assays, which typically require microgram quantities of input RNA per reaction. Moreover, comparison of in-gel eGFP fluorescence from cap-dependent (Cap1-eGFP) and IRES-dependent (IRES-eGFP) translation in rabbit reticulocyte lysate revealed that the Cap1 construct produced a strong fluorescent band, whereas IRES-driven constructs were barely detectable under identical conditions (Figure 3.9), confirming the inherently low translational efficiency of the IRES-dependent system. These two limitations, low ligation yield and low translational output, motivated the exploration of an alternative ligation strategy: chemically capping a short Nm-modified oligoribonucleotide and ligating it directly to an eGFP IVT body via a single splint junction, thereby omitting the IRES element altogether and reducing the assembly to a two-fragment reaction with substantially improved yield. This approach could also hold potential for modification mapping workflows

that require 5'-capped RNA substrates. A capping strategy compatible with the splint ligation workflow was therefore required.

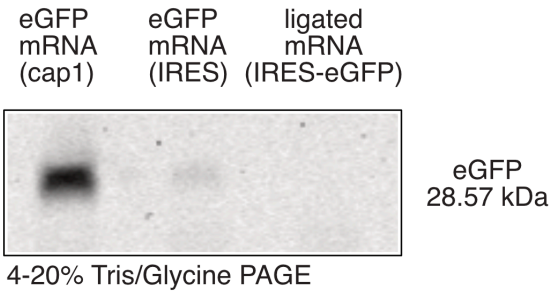


Figure 3.9 In-gel eGFP fluorescence of *in vitro* translated mRNAs. Cap1-eGFP mRNA produces a strong fluorescent signal at the expected molecular weight (28.57 kDa), whereas IRES-eGFP mRNA and splint-ligated IRES-eGFP mRNA are barely detectable, demonstrating the substantially higher translational efficiency of cap-dependent over IRES-dependent translation. Rabbit reticulocyte lysate, 4–20% Tris-glycine SDS-PAGE, fluorescence detection at 488 nm excitation.

Incorporating a 5' cap structure into the modified oligoribonucleotide required careful consideration of substrate compatibility. The synthetic oligoribonucleotide is supplied with a 5'-hydroxyl terminus and can be phosphorylated to a 5'-monophosphate by T4 polynucleotide kinase (T4 PNK), but not to the 5'-triphosphate required by enzymatic capping systems such as the Vaccinia capping enzyme. Co-transcriptional capping of the IVT fragment using ARCA was also explored, but this approach proved technically challenging. The short template (~31 nt) yielded predominantly unspecific transcripts with commercial T7 RNA polymerases, and ARCA incorporation further reduced IVT yield. Moreover, the 5'–5' triphosphate cap structure generated by ARCA would render the capped fragment incompatible with T4 RNA Ligase 2 at the ligation junction. A literature search identified m7GDP-imidazolide (m7GDP-Im) as a chemical cap donor that uniquely accepts 5'-monophosphate RNA substrates (Abe et al., 2022). The phosphorimidazolide group serves as a leaving group, facilitating nucleophilic attack by the 5'-phosphate of the RNA to form the characteristic 5'–5' triphosphate bridge of the cap structure (Sawai et al., 1999). The reaction proceeds in the presence of $CaCl_2$ as a divalent metal catalyst and 1-methylimidazole as a nucleophilic cocatalyst.

The 5′-phosphorylated oligoribonucleotide was chemically capped using m7GDP-imidazolide as cap donor, purified by denaturing PAGE (where the m^7G cap confers a +1 nt mobility shift), and visualized by Cy5 scanning or UV shadowing.

Initial capping experiments without divalent metal supplementation yielded approximately 50% capping efficiency as assessed by ImageJ densitometry on preparative gels (Figure 3.10A). Systematic optimization identified three critical parameters:

1. **Divalent metal catalyst.** Addition of $CaCl_2$ (10 mM final) to the reaction mixture containing 2 M 1-methylimidazole improved band resolution. Further titration (1, 2.5, 5, 10 mM $CaCl_2$) showed that 10 mM produced the sharpest gel bands. NaCl concentrations above 5 mM were not tolerated by the reaction, and lyophilization resulted in RNA loss.
2. **Sample preparation by SpeedVac drying.** Drying the phosphorylated oligonucleotide pellet by SpeedVac centrifugation prior to resuspension in the capping mixture increased efficiency to approximately 85%, as assessed by densitometric comparison of capped and uncapped band intensities on the preparative gel (ImageJ quantification; Figure 3.10A). This improvement is attributed to more complete removal of residual water, which competes with RNA for the phosphorimidazolide electrophile.
3. **Reaction scale.** Scaling the reaction from 100 pmol to 1 nmol reduced efficiency to approximately 20%, likely because the larger pellet did not fully dissolve in the DMSO-based capping mixture. Reactions were therefore routinely performed at 100 pmol scale and pooled when larger quantities were needed.

Product identity was confirmed by LC-MS nucleoside analysis performed in collaboration with Martina Kramer. However, the LC-MS-derived capping efficiency (50 ± 10%) was substantially lower than the gel-based estimate (~85%). This discrepancy is attributable to known limitations of LC-MS for cap quantification. The cap dinucleotide (m7GpppN) has markedly different ionization properties from standard nucleosides, and residual free m7GTP from excess capping reagent co-purifies with the RNA despite cleanup, inflating the m^7G signal (Beverly et al., 2016). The gel-based quantification, which directly resolves capped from uncapped intact oligonucleotides, is considered the more reliable estimate in this context.

Following optimization, the chemical capping protocol was applied to the six 2'-O-methylated oligoribonucleotides used for construct assembly (Section 3.1.1). Capping efficiencies ranged from 45% to 76% as assessed by gel densitometry (Figure 3.10B), with Am- and Um-containing oligos consistently achieving higher efficiencies (≥73%) than Gm-containing variants (≤62%). The position and identity of the 2'-O-methyl modification thus had a modest but reproducible effect on capping yield, possibly reflecting differences in steric accessibility of the 5'-phosphate.

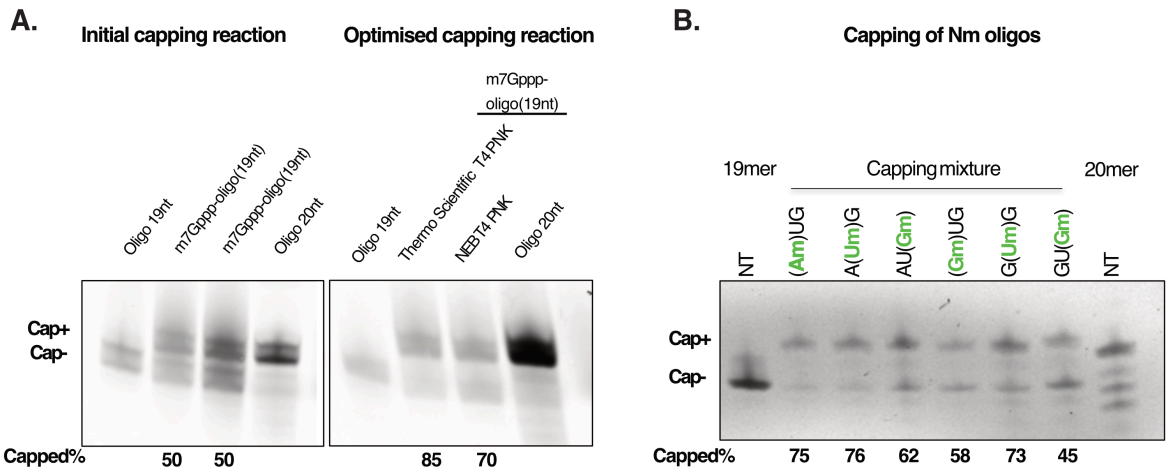


Figure 3.10 Chemical capping of 5'-phosphorylated oligoribonucleotides with m^{7}GDP-imidazolide. 20% denaturing PAGE, Cy5 scanning. (A) Left: initial capping reaction without SpeedVac drying, approximately 50% efficiency (ImageJ densitometry). The capped product (m^{7}Gppp-oligo, 19 nt) migrates with reduced electrophoretic mobility compared to the uncapped oligonucleotide. Right: optimized conditions yielding approximately 85% and 70% capping efficiency. **(B)** Application of the optimized protocol to six 2'-O-methylated oligoribonucleotides. Capping efficiencies ranged from 45% (GU(Gm)) to 76% (A(Um)G). Modified nucleotide positions are indicated in green. NT: no treatment (uncapped 19-mer and 20-mer migration references).

Following capping, the reaction mixture contains both capped and uncapped oligoribonucleotides. While gel purification can separate these species, the throughput and yield of preparative gel elution are limited. An alternative enzymatic purification strategy was therefore explored using XRN-1, a processive 5'→3' exoribonuclease that specifically degrades RNA with a 5'-monophosphate terminus while leaving 5'-capped RNA intact (Stevens, 2001).

XRN-1 dynamic range experiments confirmed that the enzyme efficiently degrades 5'-phosphorylated RNA in a dose-dependent manner, with complete digestion at 0.25 U per 20 pmol substrate using the manufacturer's protocol. However, when applied to the capping reaction mixture (250 pmol capped + uncapped RNA), the results were ambiguous. The final yield (185 pmol) exceeded the expected amount of capped RNA (~150 pmol based on 75% capping efficiency), and additional bands consistent with degradation intermediates or m^{7}GTP contamination were observed on gel analysis. Subsequent optimization attempts, including testing XRN-1 on long IVT products bearing 5'-GMP termini and varying incubation times, did not conclusively resolve whether the enzyme could reliably purify capped from uncapped species in this particular application. Although the capping protocol was developed primarily for translation assays, it also establishes a preparative foundation for nanopore-based modification detection workflows that require capped RNA substrates as defined sequencing inputs.

In summary, Section 3.1 surveyed different ligation strategies for producing site-specifically modified RNAs. Long splint ligation proved impractical due to low yields after gel purification,

whereas short 99-nucleotide constructs achieved at least three-fold higher ligation efficiency and were confirmed to be fully compatible with nanopore DRS (Section 3.1.1.4). A chemical capping protocol was developed and optimized, establishing a preparative tool with potential application in modification detection workflows that require capped RNA inputs (Section 3.1.2.3). Applied to *in vitro* translation, the capped constructs confirmed position-dependent effects of Nm on translational output (Section 3.1.2.2). Critically, the RiboMethSeq analysis revealed systematic false-positive signals indistinguishable from genuine modification (Section 3.1.1.3), underscoring the limitations of indirect biochemical readouts and motivating the transition to a nanopore-based detection strategy. Of the ligation approaches explored, the short three-fragment splint ligation method was ultimately adopted for generating the synthetic ground-truth training data used in the following nanopore DRS workflow, as the central placement of the modification site shields it from 5'- and 3'-end coverage artifacts. The concurrent development of ModiDeC (Alagna et al., 2025) and the identification of a candidate m^{5}C site on dengue virus genomic RNA by the Ruggieri Group (Wu et al., 2026) provided both the classifier framework and a biologically relevant test case. The following sections describe the resulting calibration effort.

3.2 ModiCal: Calibration-Based m^{5}C Detection by Nanopore DRS

Building on the ligation strategies established in Section 3.1, this section describes the development of the ModiCal workflow, a calibration-based approach for m^{5}C detection by nanopore DRS.

3.2.1 Rationale for Model System Selection

The development and validation of the ModiCal workflow required a model system offering well-characterized modification sites with established enzyme dependencies and prior validation by orthogonal methods. The m^{5}C2278 site in Saccharomyces cerevisiae 25S rRNA was selected as the initial target based on several criteria. Near-stoichiometric modification in wild-type cells, selective loss upon deletion of the cognate methyltransferase Rcm1 (NSUN5 homolog), and extensive prior characterization by bisulfite sequencing and mass spectrometry (Schosserer et al., 2015; Sharma et al., 2013).

A second m^{5}C site on the same rRNA molecule, C2870, modified by Nop2 (NSUN1 homolog), was deliberately excluded from initial training to serve as an internal control. This design enabled assessment of whether calibration on one site generalizes to other conserved positions,

or whether explicit inclusion in training is required for reliable detection. The following sections present the results of applying the ModiCal workflow, beginning with synthetic ground truth generation and baseline model establishment.

3.2.2 Establishing Ground Truth: Synthetic RNA Design for Baseline Training

A fundamental requirement for supervised deep learning is access to labeled training data with known ground truth. In the context of RNA modification detection, this presents a considerable challenge. Biological samples contain mixtures of modified and unmodified molecules at unknown ratios, and orthogonal methods rarely provide the single-molecule resolution needed for read-level labeling. To establish well-defined ground truth for m^{5}C detection, we therefore adopted a synthetic approach based on the short three-fragment splint ligation strategy established in Section 3.1.

3.2.2.1 Synthesis of Point-Modified Synthetic RNAs

Applying the splint ligation approach described in Section 3.1.1, we designed 99-nucleotide synthetic constructs that precisely recapitulated the native *S. cerevisiae* 25S rRNA sequence context surrounding position C2278 (corresponding to positions 2228–2327 in the full-length rRNA). Three oligoribonucleotides, a 5′ fragment, a central fragment containing the target position, and a 3′ fragment, were assembled via a DNA splint as before (Figure 3.11A–B).

For the modified construct, the central oligonucleotide (19 nucleotides) carried a 5-methylcytidine at the position corresponding to C2278 (position 50 within the 99-mer), while for the unmodified construct, the central oligonucleotide contained a canonical cytidine at the same position. Both constructs were otherwise sequence-identical, ensuring that any differences in nanopore signal could be attributed solely to the presence or absence of the methyl group.

Following phosphorylation, annealing, and overnight ligation, the products were purified by denaturing PAGE to remove incompletely ligated intermediates and unreacted oligonucleotides. Gel analysis confirmed successful assembly of both modified and unmodified full-length constructs at comparable yields (Figure 3.11C). The purified synthetic RNAs were polyadenylated and subjected to nanopore DRS using the SQK-RNA004 kit, yielding greater than 10 million reads per construct.

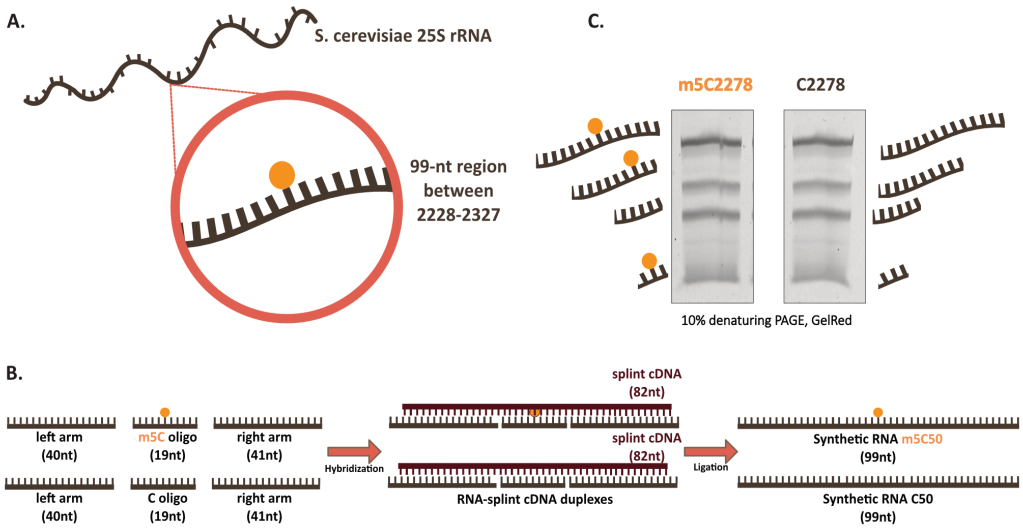


Figure 3.11 Synthetic ground-truth RNAs for baseline training. (A) *S. cerevisiae* 25S rRNA region (positions 2228–2327) containing the m^{5}C2278 site. **(B)** Splint-ligation strategy: three oligoribonucleotides annealed to a DNA splint and ligated to form the 99-nt construct. **(C)** Denaturing PAGE (10%) confirming assembly of modified and unmodified constructs.

3.2.2.2 Signal Curation and Feature Extraction

Raw nanopore signals were processed using the ModiDeC curation workflow to extract training-ready feature sets. For each sequencing read aligned to the synthetic construct reference, signal chunks centered on the target cytidine were extracted and normalized. Each chunk comprised the ionic current trace spanning five consecutive k-mers (forming a 9-mer window centered on the modification site), along with corresponding dwell times. These features were supplemented with one-hot encoded reference sequence information to provide positional context. The process is further explained in Alagna et al. (2025).

Curated signal chunks were packaged into NumPy archive (.npz) files, with 512 chunks per file serving as the basic training unit. From the initial pool of greater than 10 million reads per construct, only 100,000 reads (approximately 1%) were sampled for curation, yielding more than 1,000 .npz files per class (modified and unmodified). This deliberate subsampling ensured that training data remained manageable while still providing extensive coverage of the signal distribution.

3.2.2.3 Quantifying Training Data: Signal Chunks versus Reads

A critical but often overlooked consideration in training neural networks for nanopore modification detection is the choice of unit for quantifying training data. Raw nanopore reads vary substantially in length depending on transcript size and translocation completeness, and a single read may contain multiple modification-relevant positions or none at all. This heterogeneity means that read counts provide an unreliable measure of effective training data.

Two datasets with identical read numbers may differ substantially in the number of informative signal examples they contain for a given modification site.

To address this, ModiDeC and similar frameworks segment raw signals into fixed-length chunks centered on positions of interest, extracting normalized current, dwell time, and sequence context features for each chunk independently. These chunks are then packaged into NumPy archive (.npz) files, with a fixed number of chunks per file (512 in the ModiDeC implementation). This chunking strategy offers several advantages. It standardizes the input dimensionality for neural network training, ensures that each training example corresponds to a single modification-relevant position, and decouples training data quantity from read length variability.

Reporting training data size in terms of .npz files (or equivalently, chunk counts) rather than read counts therefore provides a more objective and reproducible metric. A dataset of 800 .npz files represents approximately 410,000 signal chunks, each centered on a defined position with consistent feature extraction, regardless of whether those chunks originated from 100,000 short reads or 10,000 long reads. This standardization proved essential during the optimization experiments described below, where we observed that read-based sampling could produce highly variable training sets depending on transcript coverage patterns. Throughout this thesis, training data quantities are therefore reported in .npz file counts to ensure comparability across experiments and facilitate reproducibility.

3.2.2.4 Optimization of Training Dataset Size

Before proceeding with baseline model training, we systematically evaluated how dataset size affects model learning and classification accuracy. Using the curated .npz files from the splint-ligated constructs, we trained a series of ModiDeC models with increasing numbers of training files per class (200, 400, 600, 800, and 1600 .npz per class) while holding all hyperparameters constant (batch size 128, k-mer model 9, four training epochs) in agreement with the original ModiDeC work.

Model performance was assessed by validation accuracy across training epochs and by a suite of classification metrics including ROC-AUC, AUPRC, accuracy, precision, and recall (Figure 3.12). The results revealed a clear pattern. Performance improved progressively as training set size increased from 200 to 800 .npz per class, at which point metrics plateaued. The 800 .npz configuration yielded the fastest and most stable convergence, achieving validation accuracy exceeding 0.95 within four epochs and producing well-balanced precision and recall. Larger training sets (1600 .npz per class) provided no additional gain in any metric, suggesting that 800

.npz per class represents a practical optimum that balances training efficiency with classification performance.

Based on these findings, all subsequent baseline training was performed using 800 .npz files per class for training (1,600 files total) and 200 .npz files per class for validation (400 files total). This standardized configuration ensured reproducibility across experiments while providing sufficient data for robust model learning.

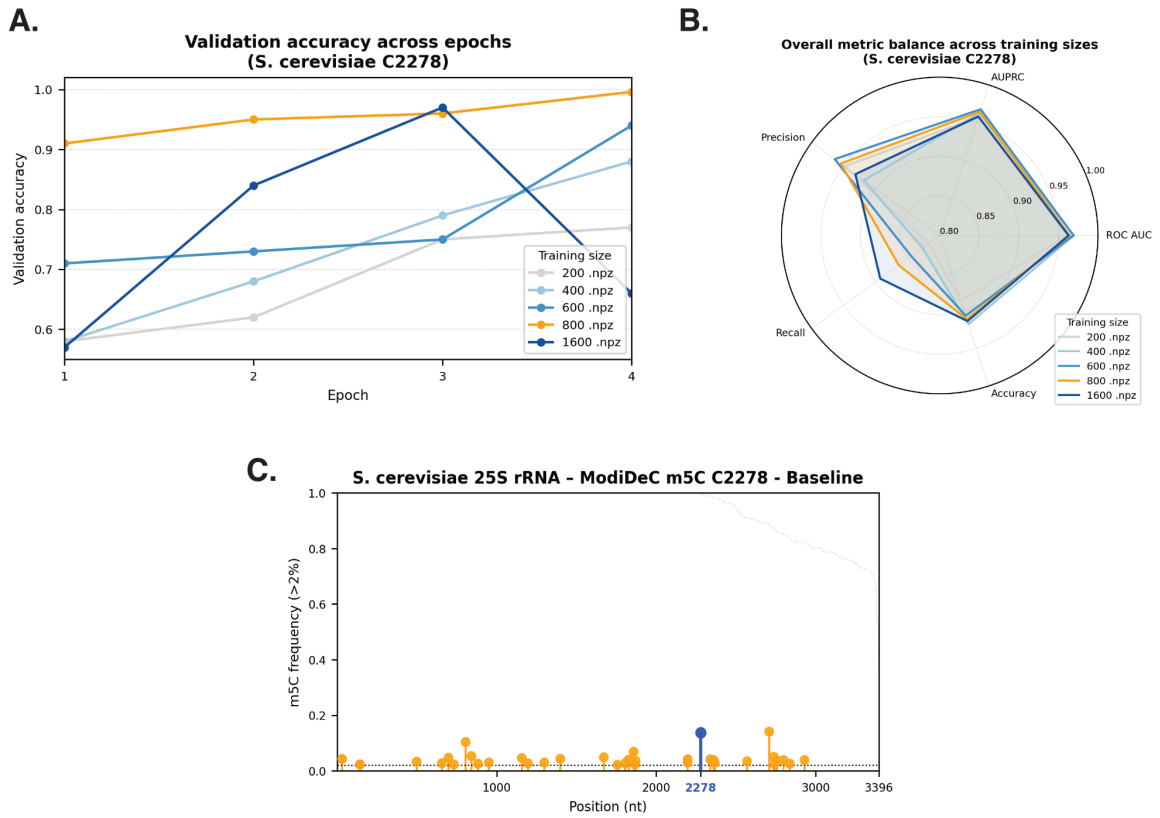


Figure 3.12 Training-size dependence of baseline model performance. **(A)** Validation accuracy across four training epochs for baseline models trained with 200–1600 .npz files per class. **(B)** Radar plot summarizing ROC-AUC, AUPRC, accuracy, precision, and recall for each training size. Optimal performance was achieved at 800 .npz per class, with no additional benefit from larger training sets. **(C)** Baseline m^{5}C prediction profile on native *S. cerevisiae* 25S rRNA. Each dot represents a cytidine position with predicted m^{5}C frequency exceeding 2%. The true-positive site C2278 is highlighted, alongside multiple false-positive calls across the transcript.

3.2.2.5 Baseline Model Performance on Native Yeast 25S rRNA

With the optimized training configuration established, baseline m^{5}C model was trained and applied to native *S. cerevisiae* 25S rRNA. For this analysis, we utilized a publicly available yeast total RNA nanopore DRS dataset (Watson et al., 2025), which provided independent biological material not used in training. The baseline model was run with the ModiDeC analysis workflow using default RNA004 settings, and per-position m^{5}C frequency was computed as the proportion of reads classified as "modified" at each cytidine.

The baseline model successfully detected the true-positive site at m^{5}C2278, confirming that the synthetic ground-truth training captured features sufficient for recognizing the modification in its native biological context (Figure 3.12C). However, two important limitations emerged. First, the predicted modification frequency at m^{5}C2278 was lower than expected (less than 20%), whereas bisulfite sequencing consistently reports near-stoichiometric modification (approximately 80%) at this position in wild-type cells (Schosserer et al., 2015). This discrepancy suggested that while the model could identify the site, it was not yet providing quantitatively accurate stoichiometry estimates.

Second, and more critically, the baseline model produced widespread false-positive predictions across the 25S rRNA sequence. A total of 35 additional cytidine positions were classified as m^{5}C-positive at frequencies exceeding the 2% detection threshold, with some false-positive sites reaching apparent modification frequencies comparable to the true-positive signal at m^{5}C2278 (Figure 3.12C). This off-target calling was not randomly distributed; certain sequence contexts appeared more prone to misclassification than others, suggesting that the model had learned features that generalized beyond the intended modification signal.

3.2.2.6 The C2870 Benchmark: Testing Generalization Limits

A particularly informative observation emerged when the baseline model's predictions were examined at the second known m^{5}C site on 25S rRNA, position C2870. Despite being modified by a different methyltransferase (Nop2/NSUN1) and located in a distinct sequence context, C2870 is also near-stoichiometrically methylated in wild-type yeast cells and exhibits enzyme-dependent loss in nop2Δ strains (Sharma et al., 2013; Bourgeois et al., 2015).

The baseline model failed to detect m^{5}C2870, predicting modification frequencies at this position indistinguishable from background (less than 2%). This negative result was significant because it demonstrated that the representation learned from C2278-specific training data does not automatically generalize to other conserved m^{5}C sites, even within the same RNA molecule. The finding raised a fundamental question. Would further calibration allow the model to eventually recognize m^{5}C2870, or would explicit inclusion in training be required? This question is addressed in Section 3.2.4.

3.2.3 Iterative Calibration: From Widespread Overcalling to Site-Specific Precision

The baseline model's performance revealed a fundamental tension between sensitivity and specificity. While the network successfully learned to recognize m^{5}C2278, it simultaneously produced widespread false-positive predictions that would undermine any attempt at biological

interpretation. This section describes the development of an iterative calibration strategy that systematically eliminates false positives while preserving, and ultimately enhancing, detection at the true modification site.

3.2.3.1 The Failure of Balanced Retraining

Before developing the full calibration workflow, it was first explored whether simple retraining strategies could correct individual false-positive sites. Signal chunks from both the modified and unmodified training pools overlapping with the most recurrent false-positive site were curated in equal numbers, and the model was retrained with these balanced additions. However, this balanced retraining failed to alter the predicted frequency at the targeted site (Figure 3.13). The outcome is mechanistically consistent with the nature of false-positive calls. A false positive reflects an existing imbalance in the network's decision boundary where unmodified signals at that sequence context are insufficiently distinguished from modified ones, and adding equal numbers of both classes preserves rather than corrects this imbalance.

This failure prompted a revised hypothesis. Because false-positive predictions arise from under-representation of unmodified signal characteristics at specific sequence contexts, correction requires asymmetric retraining that supplements the training pool exclusively with unmodified examples from the misclassified positions. By adding only negative examples, we would shift the decision boundary away from "modified" classification at those contexts while leaving the learned representation of true modification signals unchanged.

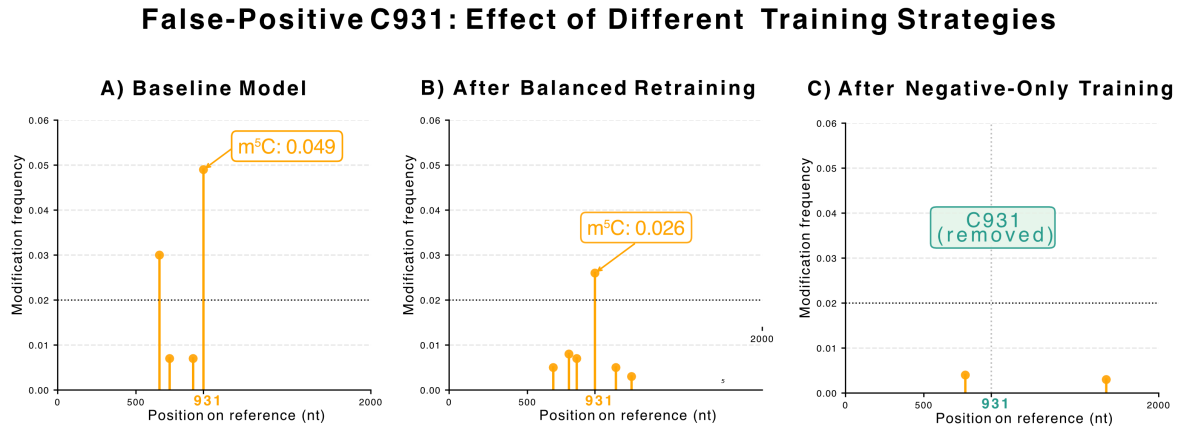


Figure 3.13 Targeted retraining strategies for false-positive elimination at C931. (A) Baseline model shows FP at 4.9%. (B) Balanced retraining with equal m^5C/C examples reduces but does not eliminate FP (2.6%). (C) Negative-only training successfully removes the FP. Dashed line indicates 2% detection threshold.

3.2.3.2 Proof of Concept: Single False-Positive Site Correction

To test the asymmetric retraining hypothesis, we repeated the experiment using only unmodified signal chunks from the targeted false-positive site, without adding corresponding modified examples. The model was retrained and re-applied to native 25S rRNA. This approach proved successful. The targeted false-positive dropped below the detection threshold, while remaining false-positive sites showed either unchanged or reduced frequencies. Importantly, the true-positive signal at C2278 remained unaffected, confirming that the negative-only retraining strategy selectively suppressed false calls without compromising genuine modification detection (Figure 3.13).

However, the practical implications of this finding presented a challenge. The baseline model had identified 35 false-positive sites across the 25S rRNA sequence. Generating individual splint-ligated constructs for each of these positions would be prohibitively costly and laborious, requiring synthesis of more than 70 additional oligonucleotides (modified and unmodified central fragments for each site) plus corresponding splints. A more scalable approach was needed.

3.2.3.3 Development of IVT-Based Bulk Calibration (BulkFP Model)

To address the scalability limitation, we reasoned that full-length IVT RNA could serve as a comprehensive source of unmodified signal for all cytidine positions simultaneously. Because IVT RNA is synthesized enzymatically using only canonical nucleotide triphosphates, it contains no m^5C modifications at any position, providing a well-defined negative reference across the entire sequence.

We generated full-length 25S rRNA by IVT from a linearized plasmid template carrying the complete *S. cerevisiae* 25S rRNA sequence under T7 promoter control. The IVT product was purified, polyadenylated, and subjected to nanopore DRS, yielding approximately 200,000 reads. Signal chunks were then curated from all 35 false-positive positions identified in the baseline analysis, sampling from 5,000 reads to produce approximately 700 .npz files containing exclusively unmodified signals.

Because unmodified signals from all false-positive positions were introduced into the training pool simultaneously, that is, in bulk, we termed the resulting model the "Bulk False Positive" or BulkFP model. The training pool composition after this step was: 800 .npz modified (splint-ligated m^5C), 800 .npz unmodified (splint-ligated canonical C), and 600 .npz IVT-derived unmodified signals from all false-positive positions. Training was performed with identical hyperparameters to the baseline model (batch size 128, k-mer model 9, four epochs).

3.2.3.4 BulkFP Model Performance

Application of the BulkFP model to native *S. cerevisiae* 25S rRNA produced a markedly improved modification profile (Figure 3.14A). The number of false-positive sites decreased from 35 to fewer than 5, representing greater than 90% reduction in off-target calling. Simultaneously, the predicted modification frequency at the true-positive site C2278 increased from less than 20% to approximately 76%, approaching the near-stoichiometric levels reported by bisulfite sequencing (Schosserer et al., 2015).

At the global level, BulkFP retraining preserved overall classification performance as assessed by ROC analysis (Figure 3.15A). The area under the ROC curve (AUC) remained high (greater than 0.95), indicating that the model's discriminative capacity was maintained despite the substantial addition of negative training examples. Precision and recall metrics similarly showed improvement, with both values increasing relative to baseline (Figure 3.15D).

Despite these improvements, a small number of false-positive sites persisted after BulkFP retraining. We reasoned that this residual background likely reflected uneven coverage across the IVT transcript. Because signal chunks were curated from a finite read pool, some false-positive positions may have been under-represented due to coverage variability or position-specific mapping biases. The question therefore arose whether complete background suppression could be achieved through additional, more targeted calibration.

3.2.3.5 Iterative Single-FP Refinement (Cal1–Cal3)

To address the remaining false positives, we implemented an iterative single-site calibration strategy that differed from the BulkFP approach in two key respects. First, rather than curating signals from all false-positive positions simultaneously, each residual site was targeted individually, ensuring explicit representation regardless of coverage heterogeneity. Second, to maintain training set balance and avoid over-fitting, only one .npz file per false-positive site was added in each calibration round.

The iterative calibration procedure operated as follows:

Cal1 (Calibration Round 1): For each false-positive site remaining after BulkFP, unmodified signal chunks were curated from full-length IVT RNA using expanded read sampling (indices 0–10,000) to ensure sufficient coverage. A single .npz file per site was appended to the existing training pool. The model was retrained with unchanged hyperparameters and re-applied to native 25S rRNA. Sites that dropped below the 2% detection threshold were considered resolved; persistent false positives were noted for inclusion in the next round.

Cal2 (Calibration Round 2): For false-positive sites that persisted through Cal1, an additional .npz file per site was added to the training pool. Retraining and re-analysis were performed as before.

Cal3 (Calibration Round 3): The procedure was repeated for any remaining false positives.

Complete background suppression was achieved after three calibration rounds (Figure 3.14B–D, 3.15B). In the final Cal3 model, only the true-positive site at m^{5}C2278 exceeded the detection threshold, with a predicted modification frequency of approximately 71%, in good agreement with bisulfite sequencing data reporting greater than 80% modification at this position (Schosserer et al., 2015). All other cytidine positions in the 25S rRNA remained below the 2% calibration threshold, demonstrating that iterative calibration can achieve essentially background-free detection when applied systematically.

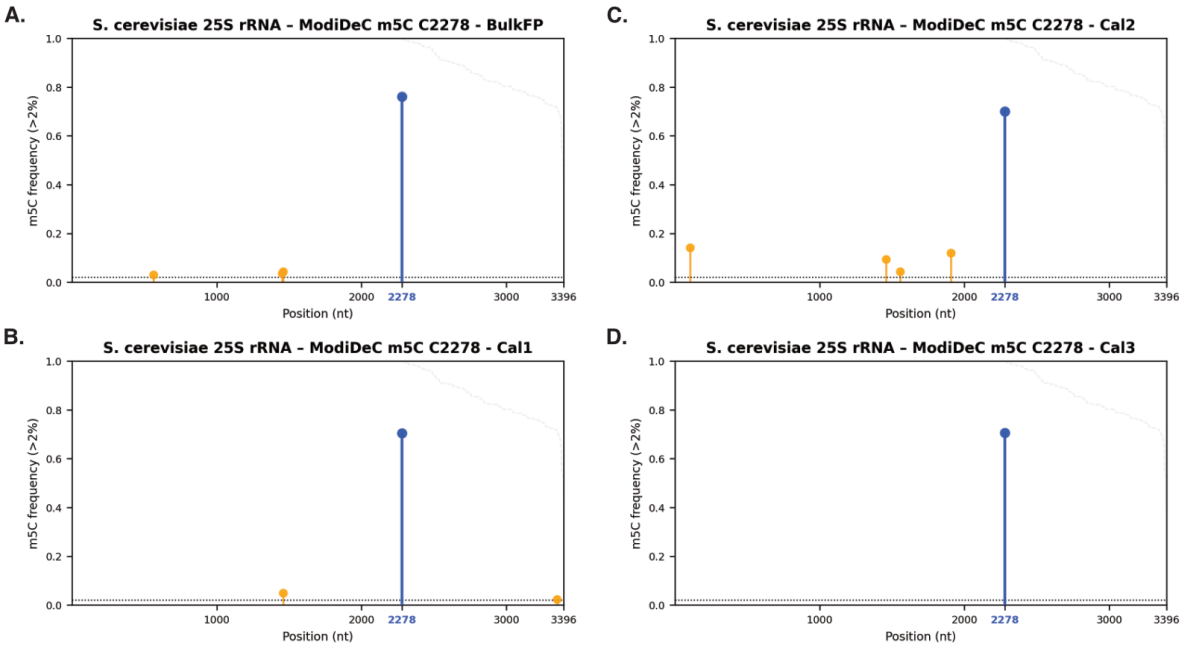


Figure 3.14 Iterative single-FP calibration progressively clears residual background in yeast 25S rRNA. (A–D) m^{5}C frequency profiles across *S. cerevisiae* 25S rRNA after BulkFP (A) and successive calibration rounds Cal1 **(B)**, Cal2 **(C)**, and Cal3 **(D)**. The predicted m^{5}C site at C2278 is indicated in blue, while residual false-positive sites are shown in orange. The dashed horizontal line denotes the 2% calling threshold. Grey traces indicate per-position read coverage.

3.2.3.6 Quantitative Validation: Synthetic Stoichiometry Titration

The calibration-induced increase in predicted modification frequency at C2278 warranted independent verification of whether the model provides quantitatively accurate stoichiometry estimates rather than merely shifting the decision threshold. To this end, an *in silico* titration was performed with defined modification stoichiometries at C2278. Nanopore DRS data from splint-ligated constructs carrying either m^{5}C or canonical cytidine at this position served as

input. Since the ground truth modification status was known for each read, reads were computationally pooled at defined ratios (0%, 5%, 10%, 15%, … 100% modified). These synthetic mixtures were then analyzed using the Cal3 model.

Predicted modification frequencies closely matched the designed input fractions across the entire stoichiometry range, with a Pearson correlation coefficient of r = 0.99 (Figure 3.15E). The relationship was approximately linear, with no evidence of systematic bias toward over- or under-estimation at any stoichiometry level. This result provided strong evidence that the calibrated model produces quantitatively reliable stoichiometry estimates, not only detecting the presence of m^5C2278 but also accurately reporting its fractional abundance.

A caveat is that the synthetic constructs used for titration represent minimal-complexity substrates; whether the observed linearity fully extends to native transcripts with complex secondary structures and neighboring modifications remains to be systematically evaluated, although the successful detection of low-stoichiometry m^5C1218 on DENV genomic RNA provides encouraging preliminary evidence.

The success of the titration experiment had important implications for the biological utility of the calibrated model. Many RNA modifications, including m^5C at various positions, occur at substoichiometric levels that vary across conditions, tissues, or developmental stages. A model that provides only binary (present/absent) information has limited value for studying such dynamic modifications. The demonstrated quantitative accuracy of the Cal3 model suggested that it could be applied to questions involving stoichiometry differences, for example, comparing modification levels between wild-type and mutant strains or across different biological conditions.

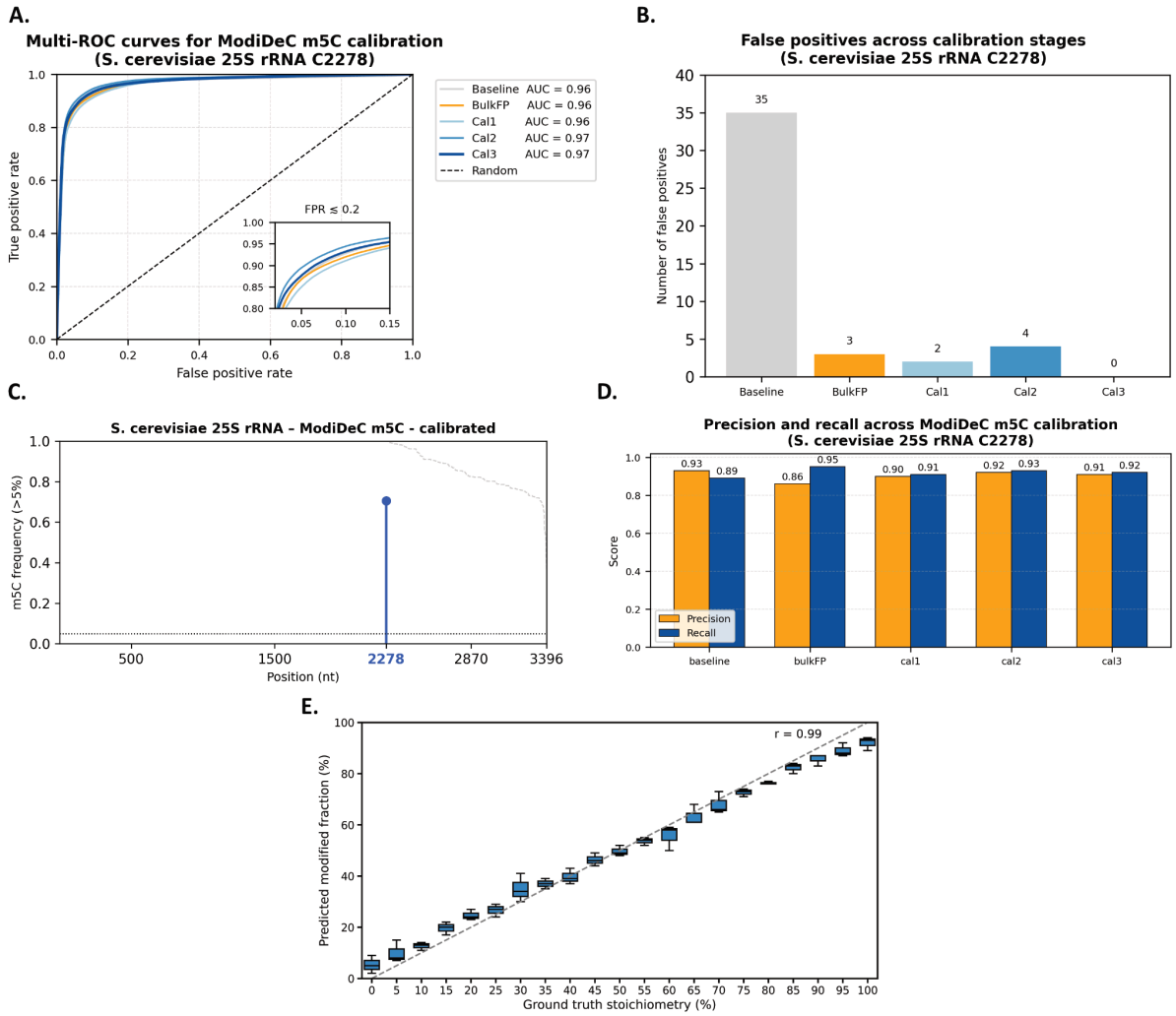


Figure 3.15 Iterative calibration eliminates false positives and preserves quantitative accuracy. **(A)** ROC curves across calibration stages, with AUC exceeding 0.98 by Cal3. **(B)** False-positive counts on native yeast 25S rRNA at each stage. **(C)** Final m^{5}C prediction profile showing detection of C2278 (~71%) with complete background suppression. **(D)** Precision and recall across calibration rounds. **(E)** Synthetic titration validation demonstrating quantitative accuracy (Pearson r = 0.99).

3.2.4 Dual-Site Calibration: Extending Detection to Multiple m^{5}C Sites on Yeast 25S rRNA

3.2.4.1 The Generalization Problem

As demonstrated in Section 3.2.2.6, the C2278-calibrated model failed to detect the second known m^{5}C site at position C2870, indicating that learned modification representations are context-specific rather than generalizable. Extending detection to both sites therefore required explicit inclusion of C2870 in training.

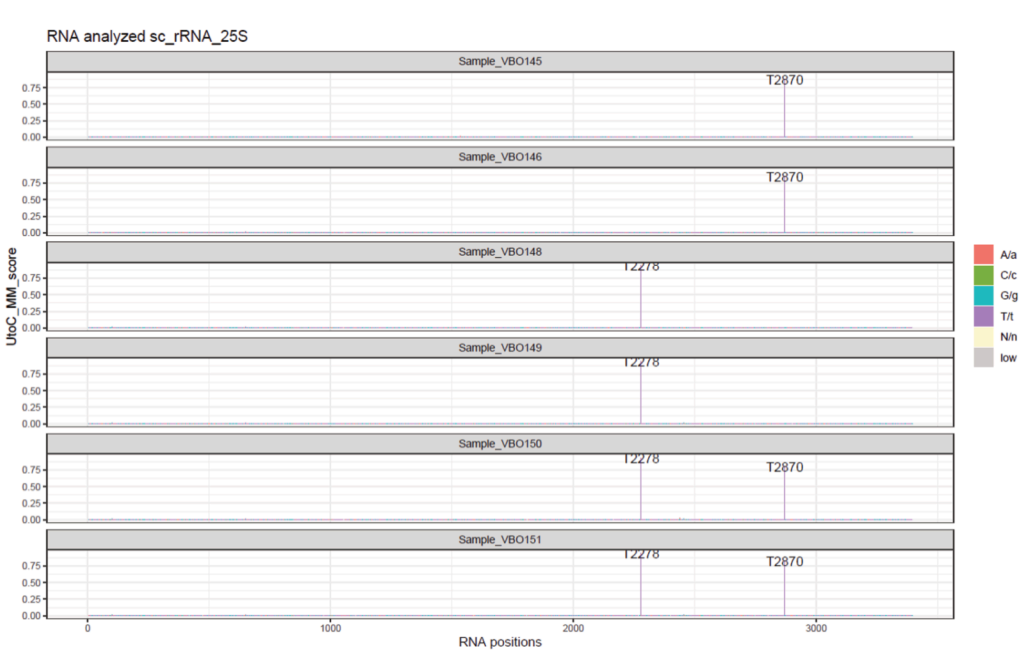
3.2.4.2 Orthogonal Validation by Bisulfite Sequencing

Before extending the ModiCal workflow to dual-site detection, we sought to establish an independent biochemical reference for our specific yeast strains and culture conditions. Total

RNA was extracted from wild-type (WT), rcm1Δ, and nop2Δ *S. cerevisiae* strains and subjected to bisulfite sequencing as described in Methods. Bisulfite treatment converts unmodified cytidines to uridines through oxidative deamination, whereas m^5C is protected from this conversion, allowing site-specific quantification based on the ratio of residual cytidine signal at each position (Squires et al., 2012).

Consistent with previous reports, wild-type 25S rRNA displayed two prominent bisulfite-resistant peaks at positions C2278 and C2870, with modification levels exceeding 80% at both sites (Figure 3.16). In the *rcm1Δ* strain, the C2278 signal dropped to baseline levels indistinguishable from fully converted cytidines, while C2870 remained at near-stoichiometric modification. Conversely, the *nop2Δ* strain showed selective loss of the C2870 signal with retention of wild-type-like modification at C2278. Importantly, all other cytidine positions in 25S rRNA remained at baseline levels across all strains, confirming that C2278 and C2870 are the only detectable m^5C sites under our experimental conditions. These bisulfite sequencing results validated the expected enzyme-dependent modification patterns and provided a quantitative reference against which to evaluate nanopore-based detection.

A.



B.

Sample	Aligned reads	Strain genotype	Referred as
VBO145	12,504,176	ΔYNL022c mat Alpha	Rcm1 knockout, Replicate 1
VBO146	14,873,434	ΔYNL022c mat A	Rcm1 knockout, Replicate 2
VBO148	14,459,538	ΔYNL061 + Nop2 DB_1B n°1	Nop2 knockout, Replicate 1
VBO149	14,451,874	ΔYNL061 + Nop2 DB_1C	Nop2 knockout, Replicate 2
VBO150	15,911,927	ΔYNL061 + Nop2 DB_1N n°3	Wild Type, Replicate 1
VBO151	15,842,563	ΔYNL061 + Nop2 DB_2C	Wild Type, Replicate 2

Figure 3.16 Bisulfite sequencing of m^5C sites on *S. cerevisiae* 25S rRNA. (A) UtoC_MM score profiles along the yeast 25S rRNA for Rcm1 knockout, Nop2 knockout, and wild-type samples. rRNA modification mapping was done using BSSeq protocol, potentially detecting bisulfite-resistant m^5C / m^4C RNA residues. The BSSeq signal shows residual non-deaminated C content in RNA (UtoC_MM_score) (see Methods). RNA sequence and numbering of positions are shown at the bottom. Color code corresponds to the nature of nucleotide in RNA. Vertical lines indicate positions C2278 and C2870. **(B)** Overview of samples, aligned read counts, strain genotypes, and experimental grouping used in this analysis.

3.2.4.3 Training Strategy for Dual-Site Detection: Combining Synthetic and Native Ground Truth

To extend the single-site model to known 25S rRNA m^5C positions, we required ground truth training data for C2870. Unlike C2278, for which splint-ligated synthetic constructs provided well-defined modified and unmodified reference signals, generating equivalent synthetic material for C2870 would require additional oligonucleotide synthesis with the attendant costs and time. We therefore adopted an alternative strategy that leveraged the availability of knockout strains. Rather than synthetic ground truth, we curated training data for C2870 directly from native nanopore reads using genetic controls.

Total RNA from wild-type and *nop2*Δ cells was sequenced by nanopore DRS, yielding approximately 7 million reads per sample with approximately 300,000 reads aligning to 25S rRNA. Signal chunks corresponding to the C2870 position were curated from both datasets. Wild-type signals served as the "modified" class (containing m^5C2870 at near-stoichiometric levels), while *nop2*Δ signals served as the "unmodified" class (lacking m^5C2870 due to enzyme deletion). These C2870-specific training files were added in balanced numbers to the existing C2278-based training pool, which consisted of synthetic modified and unmodified signals from splint-ligated constructs.

Consequently, the expanded model learned m^5C2278 from synthetic ground truth and m^5C2870 from native genetic controls. This hybrid approach illustrates a practical advantage of the ModiCal framework. When knockout strains are available, they can substitute for synthetic constructs as sources of modification-free reference signals. The expanded model was then calibrated using the same iterative workflow established in Section 3.2, incorporating unmodified IVT signals from false-positive positions to suppress background.

3.2.4.4 Validation Design: Ensuring Independence of Training and Test Data

A critical consideration when using native reads for both training and validation is the risk of data leakage. If the same reads are used for training and testing, apparent performance may be inflated by overfitting rather than genuine generalization. To avoid this confound, we implemented strict data partitioning throughout the dual-site experiments. Training data for C2870 was curated from a dedicated subset of wild-type and *nop2*Δ reads that were excluded

from all downstream analyses. The modification profiles presented in Figure 3.17 were generated exclusively from non-overlapping read sets, ensuring that detection of m^5C2870 reflects true modification status rather than recognition of previously seen signals.

Furthermore, biological replicates were processed independently to assess reproducibility. Total RNA was extracted from two independent cultures for each strain (WT, *rcm1*Δ, and *nop2*Δ), yielding six samples that were sequenced in separate library preparations. The consistency of modification profiles across biological replicates provides confidence that the observed patterns reflect genuine biological signals rather than technical artifacts or batch effects.

3.2.4.5 Enzyme-Dependent Detection in Biological Samples

Application of the calibrated dual-site model to native yeast 25S rRNA produced clean and readily interpretable modification profiles across all strains (Figure 3.17A, left panels). In wild-type cells, two clear m^5C peaks were detected: C2278 at approximately 76-78% modification frequency and C2870 at approximately 55-57%. These values are consistent with the near-stoichiometric modification reported by bisulfite sequencing, although the C2870 frequency appears somewhat lower than the C2278 frequency, possibly reflecting differences in signal accessibility or k-mer context effects between the two positions.

The enzyme-dependent patterns predicted by the genetic model were faithfully recovered. In *rcm1*Δ cells, the C2278 signal dropped to baseline levels (below 5% threshold), confirming the complete loss of Rcm1-dependent methylation. Simultaneously, C2870 remained reliably detected and showed a modest increase in apparent modification frequency (approximately 70-72%), potentially reflecting the absence of cross-talk effects or changes in rRNA processing in the knockout background. Conversely, in *nop2*Δ cells, the C2870 peak was selectively lost while C2278 was retained at wild-type-like levels (approximately 78-80%). These reciprocal patterns were consistent across biological replicates, demonstrating the reproducibility of the dual-site detection.

Critically, apart from C2278 and C2870, no other cytidines in 25S rRNA exceeded the detection threshold in any strain or replicate when analyzed with the calibrated model. The modification profiles were remarkably clean. Only the two bona fide sites emerged above background, validating the effectiveness of the calibration workflow in suppressing false-positive calls while preserving sensitivity at true modification sites. This background-free behavior extended across the entire 3,396-nucleotide rRNA sequence, indicating that the calibration successfully addressed context-dependent overcalling across diverse sequence environments.

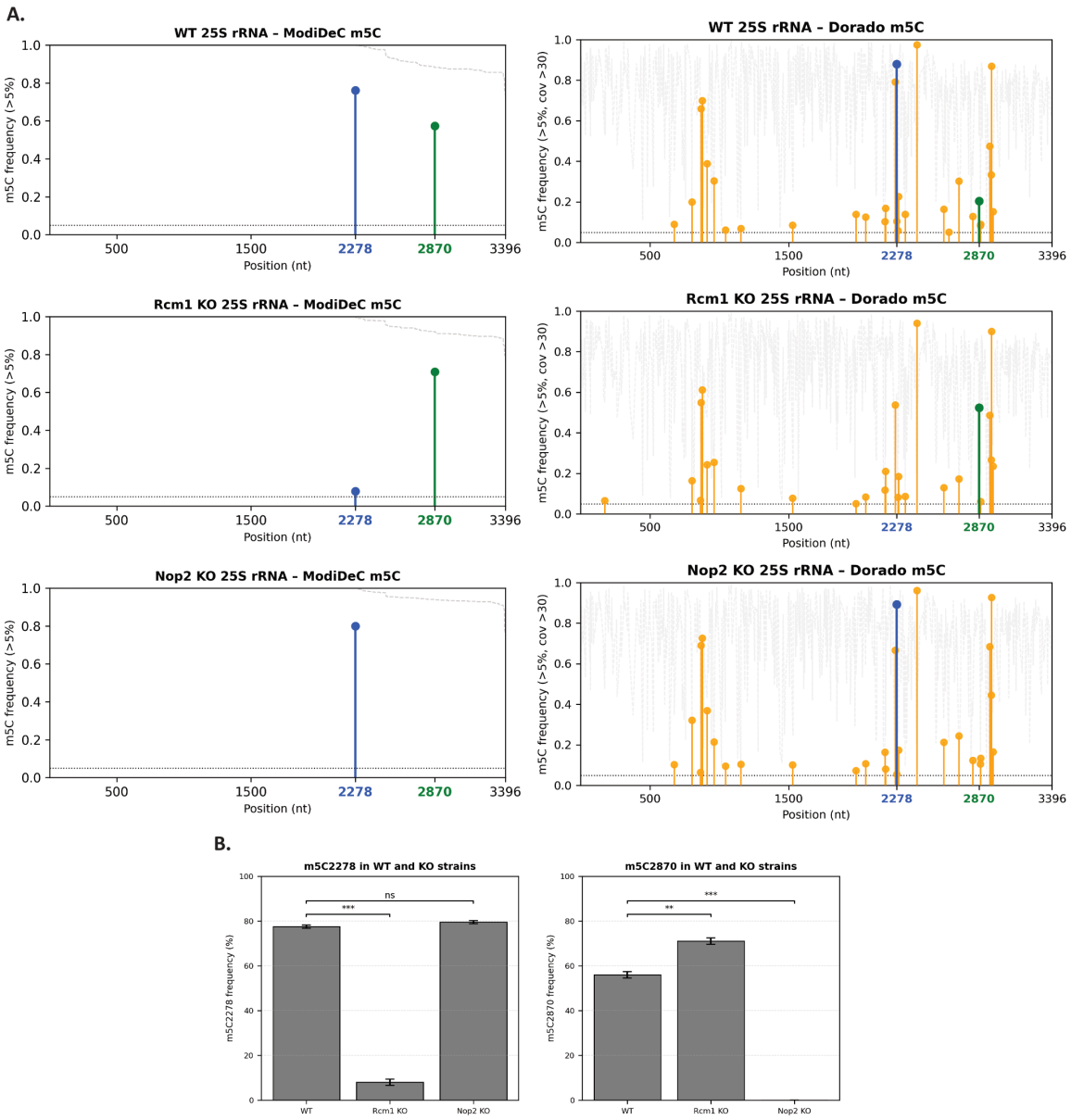


Figure 3.17 Validation of calibrated m^{5}C model in wild-type and knockout yeast strains. (A) m^{5}C profiles across *S. cerevisiae* 25S rRNA obtained with the calibrated ModiDeC model (left panels) and Dorado m^{5}C model (right panels) for wild-type (WT), *rcm1*Δ, and *nop2*Δ samples. True-positive sites are indicated: C2278 (blue) and C2870 (green). Off-target calls exceeding the detection threshold are shown in orange. Grey traces indicate per-position read coverage. The calibrated model detects only the two bona fide sites with their correct enzyme dependencies, whereas Dorado produces approximately 25 additional high confidence calls at the 0.95 probability cutoff. **(B)** Quantification of m^{5}C2278 and m^{5}C2870 modification frequencies in WT, *rcm1*Δ, and *nop2*Δ samples using the calibrated model, showing selective loss of C2278 in *rcm1*Δ and of C2870 in *nop2*Δ. Error bars represent variation between biological replicates.

3.2.4.6 Comparison with Modification-Aware Basecaller: Dorado m^{5}C Model

To contextualize the performance of the calibrated ModiCal model, we analyzed the same sequencing reads using Dorado, the modification-aware basecaller developed by ONT (Figure 3.17A, right panels). Dorado correctly detected both true-positive sites (C2278 and C2870) with their expected enzyme dependence: C2278 was present in wild-type and *nop2*Δ but absent in

*rcm1*Δ, while C2870 was present in wild-type and *rcm1*Δ but absent in *nop2*Δ. This confirms that Dorado's m^{5}C model can detect genuine modification signals under favorable conditions.

However, the Dorado output presented a substantially more complex landscape than the ModiCal profiles. At the commonly used probability cutoff of 0.95, Dorado reported approximately 25 additional high-confidence m^{5}C calls across 25S rRNA beyond the two true-positive sites. Several of these spurious calls exceeded 80% apparent modification frequency and were consistently detected across all strains, including the knockout samples where no additional m^{5}C modifications are expected. This pattern is consistent with the false-positive inflation previously reported in independent benchmarking studies (Guo et al., 2025; Luo et al., 2025).

Analysis of Dorado output at different probability thresholds revealed a trade-off between coverage and specificity (Figure 3.18). The stringent 0.95 cutoff filtered out many reads, reducing per-position coverage and creating uneven depth profiles (visible as the light grey trace in Figure 3.17A). Lowering the cutoff to 0.50 restored more uniform coverage but markedly increased false-positive calls from approximately 25 to approximately 160 sites on average. This threshold dependence illustrates a fundamental challenge with modification-aware basecallers: the user must choose between stringent thresholds that sacrifice coverage or permissive thresholds that inflate false-positive rates, with no threshold achieving the background-free behavior demonstrated by the calibrated ModiCal model.

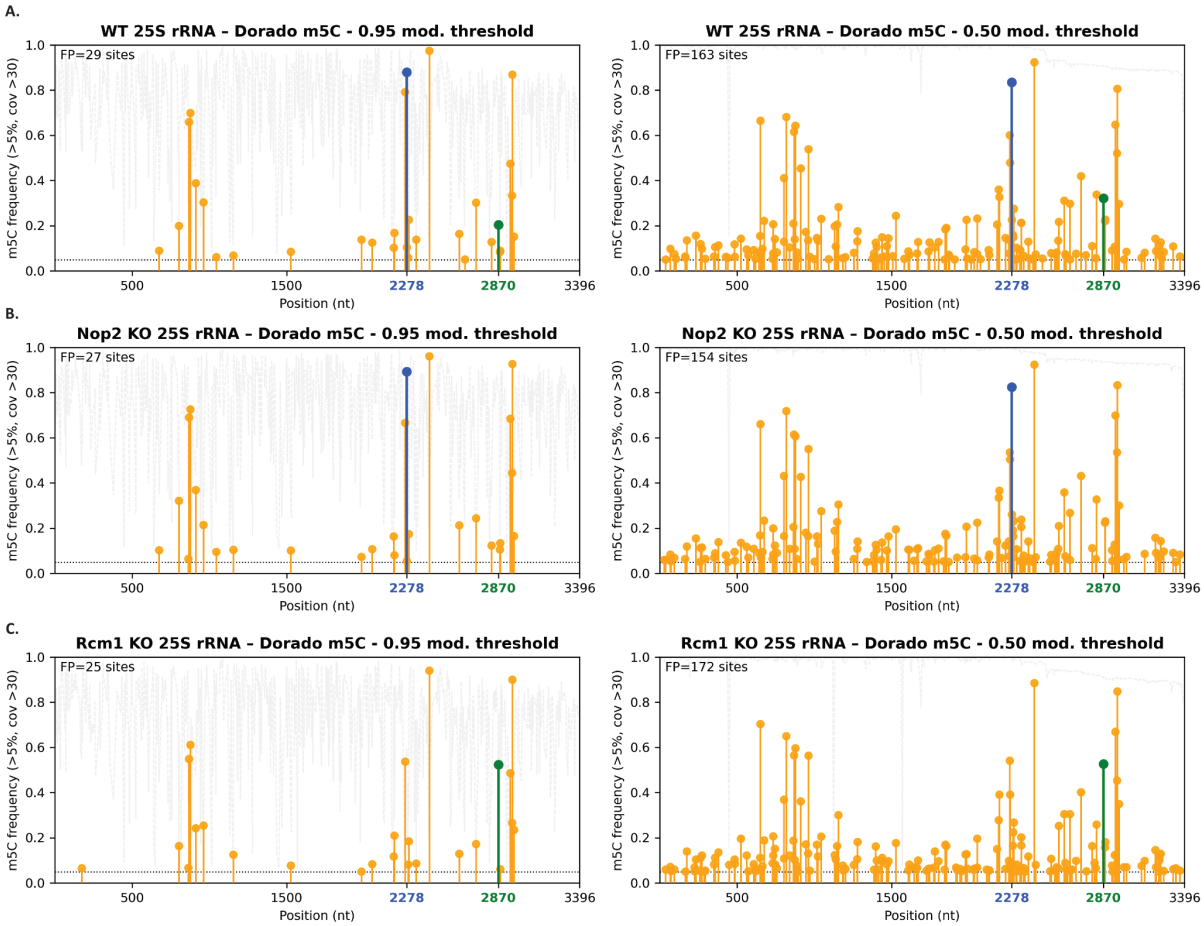


Figure 3.18 Threshold dependence of Dorado m^{5}C detection on yeast 25S rRNA. m^{5}C prediction profiles using Dorado at stringent (0.95, left) and permissive (0.50, right) probability thresholds across wild-type (A), nop2Δ (B), and rcm1Δ (C) samples. True-positive sites at C2278 (blue) and C2870 (green) are indicated. False-positive counts (orange, FP) increase substantially at the lower threshold (from 25–29 to 154–172 sites), illustrating the trade-off between coverage retention and specificity. Grey traces indicate positions below the 5% detection threshold or with coverage below 30 reads.

3.2.5 Cross-RNA Transferability: Application of the ModiCal Workflow to Dengue Virus Genomic RNA

3.2.5.1 Rationale for Testing on an Independent RNA System

The experiments described in Sections 3.2-3.4 established ModiCal as a reliable framework for high-precision m^{5}C detection on *S. cerevisiae* 25S rRNA, achieving complete background suppression while preserving enzyme-dependent detection of both characterized modification sites. Whether the calibration logic transfers to structurally and biologically distinct RNA systems or remains specific to the ribosomal RNA context in which it was developed remained to be determined. To address this question, we applied the ModiCal workflow to the genomic RNA of dengue virus (DENV), a single-stranded positive-sense RNA of approximately 10.7 kb,

nearly three times longer than 25S rRNA, with complex secondary structures distinct from ribosomal RNA architecture. The candidate m^{5}C site at position 1218, recently identified within the envelope (E) protein coding region and attributed to NSUN6, presents a particularly stringent test case: bisulfite sequencing estimated the modification stoichiometry at only 15–25%, and evidence has relied primarily on Dorado-based nanopore analysis without orthogonal single-site validation (Wu et al., 2026).

3.2.5.2 Synthetic Ground Truth Generation for DENV C1218

Following the established three-fragment splint ligation design (Section 3.2.2), we assembled 99-nt synthetic constructs recapitulating the native DENV sequence context around C1218, with either m^{5}C or canonical C at the target position. Purified products were sequenced by nanopore DRS, yielding approximately 5 million (unmodified) and 8 million (modified) reads.

As a comprehensive unmodified reference for calibration, we utilized full-length DENV genomic RNA generated by IVT. The IVT product and corresponding nanopore DRS data were obtained directly from Wu et al. (2026). Because IVT RNA contains only canonical nucleotides, it provides a well-defined negative control at all cytidine positions across the entire 10.7 kb viral genome.

3.2.5.3 Baseline Model Performance on Native DENV Genomic RNA

Signal chunks from the modified and unmodified splint-ligated constructs were curated and trained a baseline DENV model following the yeast protocol (Section 3.2.2). Given the low stoichiometry reported for m^{5}C1218, we adopted a more permissive detection threshold of 2% to retain genuine low-level modification signals while still providing some degree of background filtering.

Application of the baseline DENV model to native viral genomic RNA revealed a pattern notably reminiscent of the initial yeast results (Figure 3.19A, top panel). The model correctly identified a peak at position C1218, with an apparent modification frequency of approximately 33%, consistent with the bisulfite sequencing estimates. However, the baseline prediction also produced a dense background of miscalls distributed across the viral genome. A total of 374 cytidine positions exceeded the 2% threshold, with several false-positive sites displaying apparent modification frequencies above 80%. This widespread overcalling rendered the modification landscape effectively uninterpretable: although the true-positive site was detected, it could not be confidently distinguished from the numerous spurious calls without additional information.

3.2.5.4 Bulk False-Positive Calibration Substantially Simplifies the DENV Modification Landscape

To address the extensive false-positive background, we applied the BulkFP calibration strategy developed on yeast rRNA. Unmodified signal chunks were curated from all 374 false-positive positions using the DENV IVT transcript as the source of modification-free reference signals. These IVT-derived chunks were added to the training pool alongside the original splint-ligated modified and unmodified datasets, and the model was retrained with identical hyperparameters.

The effect of BulkFP calibration was pronounced (Figure 3.19A, middle panel). After this single retraining step, the number of false-positive sites dropped from 374 to just one residual position (Figure 3.19C), representing a greater than 99% reduction in off-target calls. Simultaneously, the true-positive signal at C1218 was preserved at approximately 31% modification frequency, indicating that the calibration successfully suppressed background without impairing sensitivity at the bona fide site. Global classification performance improved correspondingly, with AUC increasing from 0.94 at baseline to 0.99 after BulkFP calibration (Figure 3.19B), and precision shifting from widespread overcalling to balanced performance (Figure 3.19D).

3.2.5.5 Iterative Single-Site Refinement Achieves Complete Background Suppression

To eliminate the single residual false-positive site remaining after BulkFP calibration, we applied the iterative single-site refinement strategy. Unmodified signal chunks were curated specifically from the remaining false-positive position using IVT reads, and one additional .npz file was added to the training pool. This targeted approach was repeated for three calibration rounds (Cal1-Cal3) to ensure complete and stable background suppression.

Complete background suppression was achieved by the Cal3 model (Figure 3.19A, bottom panel; Figure 3.19C–D). Analysis of native DENV genomic RNA with the final calibrated model revealed a single peak at position C1218 with approximately 29% modification frequency, while all other cytidine positions across the 10.7 kb genome remained below the 2% detection threshold. The true-positive signal remained stable throughout the calibration process, with modification frequencies of 31%, 30%, and 29% at the BulkFP, Cal1-Cal2, and Cal3 stages respectively. This consistency indicates that the calibration procedure refines specificity without progressively eroding sensitivity.

The successful transfer of the ModiCal workflow to DENV genomic RNA, achieving background-free detection of a low-stoichiometry site on a 10.7 kb transcript, establishes the generalizability of the calibration framework beyond the yeast rRNA system in which it was developed.

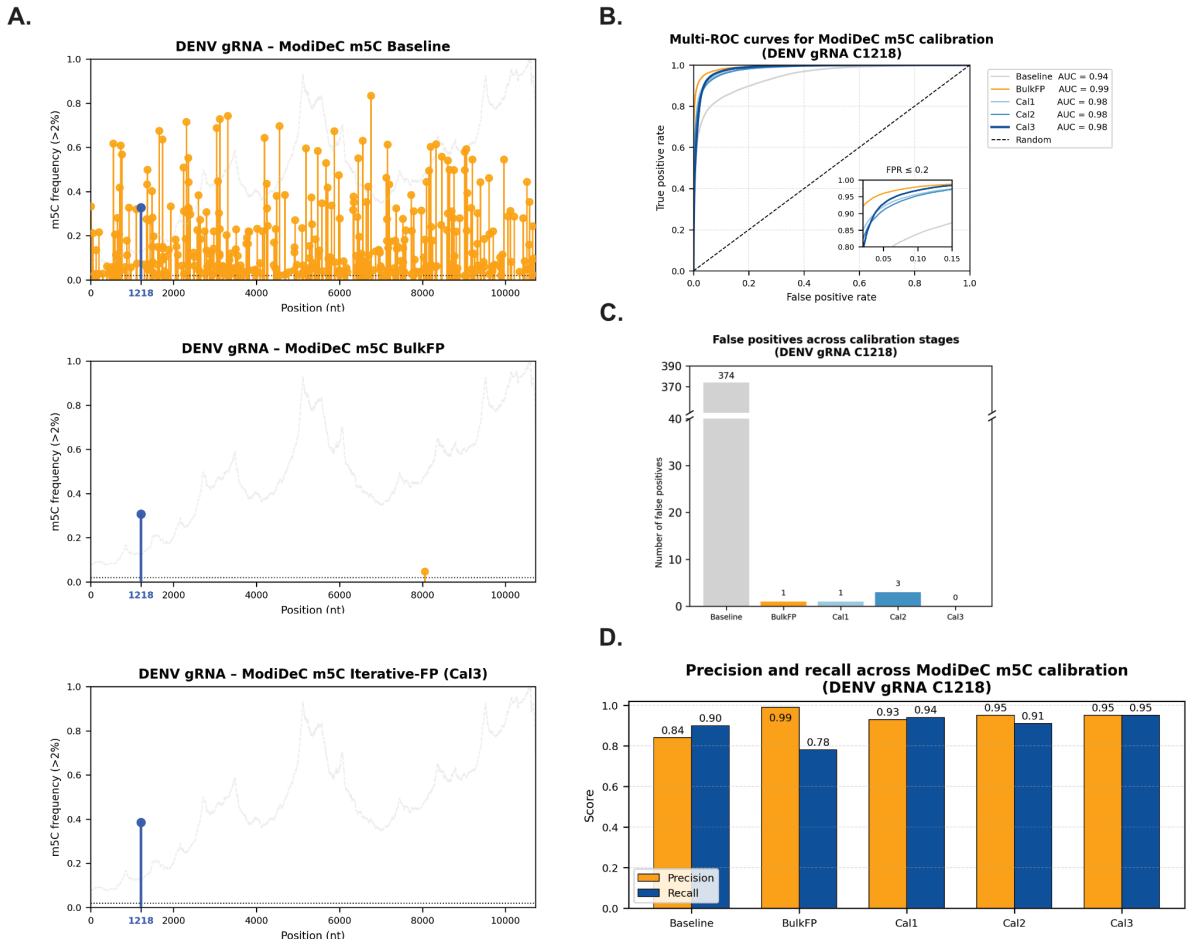


Figure 3.19 Calibration of the ModiCal workflow for Dengue virus genomic RNA. (A) m^{5}C prediction profiles across the DENV gRNA (10.7 kb) at baseline, BulkFP-calibrated, and final Cal3 stages. The true-positive site at C1218 (blue) is detected throughout calibration, while false positives (orange) decrease from 374 to 0. The final model detects C1218 at approximately 29% modification frequency with complete background suppression. (B) Multi-ROC curves showing AUC improvement from 0.94 (baseline) to 0.99 (BulkFP). (C) False-positive counts across calibration stages. (D) Precision-recall shift from baseline overcalling to balanced performance in the final model.

3.2.6 A Generalized Decision Framework for ModiCal Implementation

The calibration workflow established through the yeast rRNA and DENV experiments can be distilled into a generalizable decision framework for researchers seeking to apply ModiCal to novel RNA targets. Figure 3.20 presents this framework as a flowchart that guides users through the three calibration stages with defined decision points and stopping criteria.

The flowchart encodes the empirical thresholds identified in this work. A minimum of 800 .npz files per class for baseline training, a validation accuracy target of 0.95 to diagnose under- or overfitting, and iterative refinement until zero detectable false positives are achieved. At each decision node, the flowchart specifies whether to increase training data, add IVT-derived negative examples from false-positive positions, or proceed to single-site targeting. This

structured approach enables systematic application of the calibration logic to any RNA system where synthetic ground truth constructs and IVT reference material can be generated.

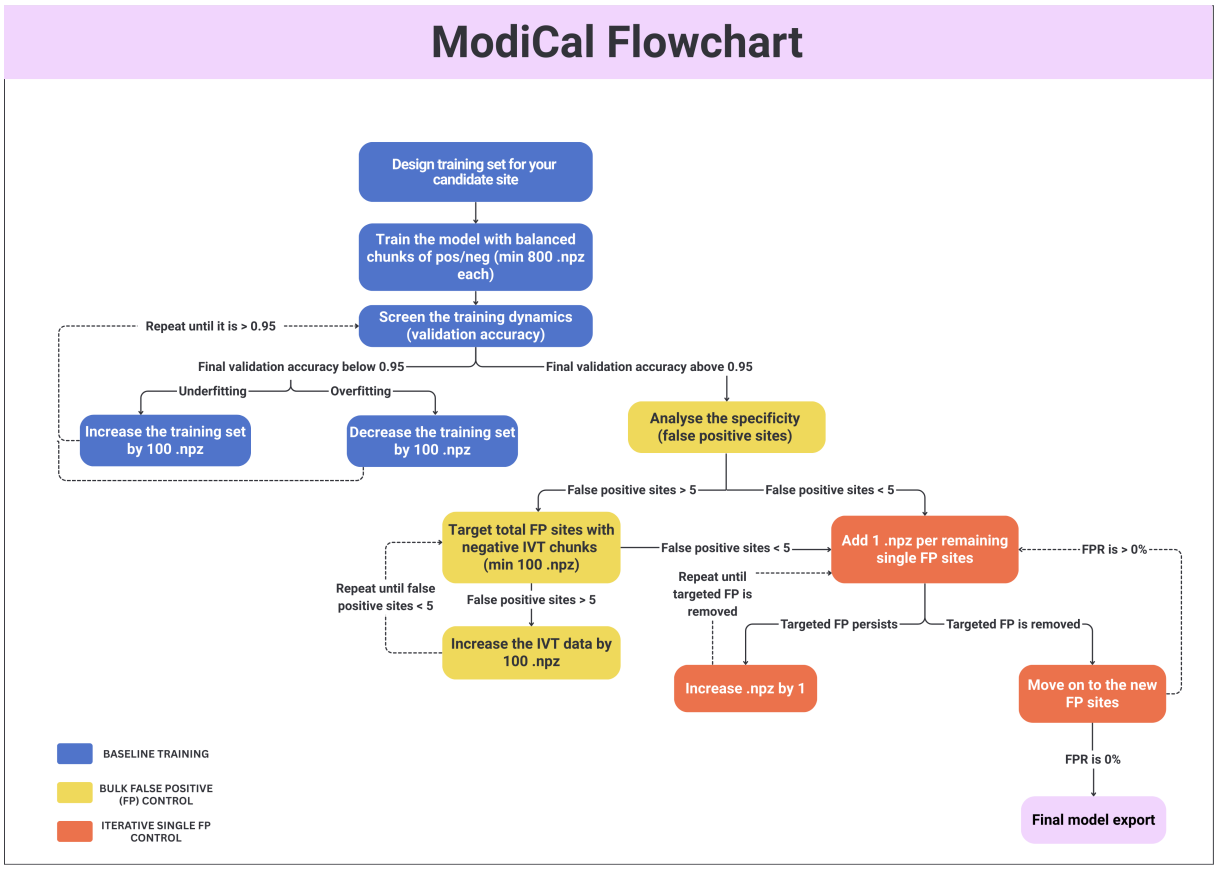


Figure 3.20 Decision flowchart for ModiCal implementation. The workflow proceeds through three stages: baseline training (blue), bulk false-positive calibration (yellow), and iterative single-site refinement (orange).

4. Conclusion and Perspectives

4.1 Summary of Ligation Strategies and Biochemical Foundations

The initial phase of this thesis set out to validate putative Nm sites on SARS-CoV-2 gRNA using RiboMethSeq analysis of site-specifically modified constructs generated by splint-mediated ligation. The candidate Nm sites on Spike and Nsp13 were demonstrated to be false positives: unmodified control RNAs produced protection signals indistinguishable from those of the modified constructs, and an additional false positive at a position where no modification had been introduced confirmed that the observed signals were technical artifacts rather than biological. These findings exposed a fundamental limitation of cleavage-based modification detection, in which sequence- and structure-dependent protection of phosphodiester bonds generates systematic false positives that cannot be resolved by threshold adjustment alone.

Alongside the RiboMethSeq work, several technical advances were made. A systematic comparison of splint ligation formats demonstrated that short constructs are substantially more efficient to produce than long ones and are fully compatible with nanopore DRS, establishing a practical pipeline for generating synthetic ground-truth training data. A position-scanning library of 2′-O-methylated mRNAs revealed that methylation at the first position of the start codon nearly completely abolishes IRES-dependent translation, although results at other positions remained inconclusive due to the limited dynamic range of the system. To support a planned transition to cap-dependent translation, an optimized chemical capping protocol using m^7GDP-imidazolide was developed, providing an efficient post-transcriptional capping method applicable to any 5′-phosphorylated RNA. Collectively, these biochemical studies established the technical foundations for working with site-specifically modified RNAs while simultaneously revealing the false-positive problem that would recur, in a different technical manifestation, in nanopore-based modification detection.

4.2 ModiCal: Calibration-Based Modification Detection

The central computational contribution of this thesis is ModiCal, a calibration-based framework for high-precision m^5C detection by nanopore DRS. ModiCal addresses the systematic overcalling that arises when neural network classifiers are applied to biological

RNA, where sequence context, secondary structure, and stoichiometric variation generate false-positive modification calls analogous to those observed in the RiboMethSeq experiments. Rather than treating these false positives as noise to be filtered post hoc, ModiCal repurposes them as diagnostic readouts that reveal where the classifier's decision boundaries are misaligned with biological reality and uses this information to iteratively refine the model.

The framework rests on three calibration principles: balanced addition of unmodified and modified ground truth data targeting the candidate site; full-length IVTs as a scalable negative reference that captures the sequence-context diversity of the target RNA; and a two-stage workflow combining bulk false-positive suppression with iterative single-site refinement. Application to yeast ribosomal RNA and subsequent transfer to dengue virus genomic RNA demonstrated that the calibration logic generalizes across structurally and biologically distinct systems. ModiCal is explicitly not a *de novo* discovery tool; it is designed to be coupled with transcriptome-wide discovery methods that nominate candidate sites, providing the calibrated, site-specific validation required before mechanistic follow-up.

4.3 Limitations and Future Perspectives

The transition from IRES-dependent to cap-dependent translation was initiated but not completed. Enzymatic purification of capped versus uncapped oligonucleotides still requires optimization, and the kinetics of XRN-1 digestion should be systematically assessed (e.g., across enzyme concentration and incubation time) to establish reliable conditions. Even if purification can be achieved, direct visualization of capped-oligo ligation to eGFP mRNA remains challenging, as the expected mobility shift on gel is subtle and often difficult to resolve. To improve detectability, a fluorescently labeled DNA splint was tested as a proxy readout, with the aim of observing a fluorescence band shift upon binding to the ligated construct; however, this approach also requires further optimization.

ModiCal depends on prior knowledge of at least one confirmed modification site for initial training, has been validated only for m^5C, and requires substantial computation time with manual intervention between calibration stages. Validation has so far been performed on transcripts up to approximately ten kilobases in length; applicability to substantially longer RNAs remains to be tested.

These limitations suggest several directions for future development. Although the current calibration workflow was developed on RNA004 chemistry, the underlying principles are chemistry-agnostic. In case successor chemistries with improved signal resolution become

available, the same calibration logic would apply and may require fewer iterative rounds owing to better discrimination between modified and unmodified current signatures. A natural next step is the extension to other modification types, particularly 2′-O-methylation, which was the starting point of this thesis and for which the splint ligation pipeline is already established. The successful calibration of two m^5C sites on a single RNA suggests that the framework could be adapted to modification-dense substrates such as tRNAs, where multiple modification types co-occur at high density; transfer learning across such targets may prove valuable in reducing the calibration effort required for each new system. Co-calibration of different modification types on the same transcript is conceptually feasible, though not presently required given ModiCal's current role as a validation tool. Beyond ModiDeC, alternative *de novo* modification prediction tools could benefit from the same calibration logic, establishing the generality of the approach across the nanopore toolkit. A frozen-baseline architecture, in which the initial model remains fixed while calibration-specific layers accumulate incrementally, could reduce computational requirements, and a community-driven calibration database for common targets would minimize redundant effort across laboratories. Workflow integration (EPI2ME) can be improved, as calibration currently requires impractical manual identification and inclusion of false-positive sites into training. Automating this iterative loop would improve time efficiency, reproducibility, and user-friendliness, moving towards a single-command calibration workflow.

4.4 Concluding Remarks

This thesis contributes both biochemical method development and computational innovation to the field of RNA modification detection. The discovery of systematic false positives in RiboMethSeq provides cautionary evidence that threshold-based calling on indirect readouts is insufficient for reliable modification mapping, a lesson that extends beyond any single method. The chemical capping protocol fills a practical gap in the toolbox for producing site-specifically modified RNAs, and the splint ligation pipeline provides a generalizable route to synthetic ground-truth standards. ModiCal, built on the conceptual foundation that false positives are informative rather than merely problematic, offers a principled path from overcalling to clean, validation-grade modification profiles. Taken together, these contributions demonstrate that the challenge of false positives in epitranscriptomics, far from being an obstacle, can be reframed as a tool for iterative methodological refinement.

5. Materials and Methods

5.1 Materials

5.1.1 Instruments

Instrument	Supplier
Thermoshaker Plus	Eppendorf (Hamburg, Germany)
Thermal cycler (PeqSTAR 96X)	Peqlab (Erlangen, Germany)
Eppendorf Centrifuge 5424 R	Eppendorf (Hamburg, Germany)
SpeedVac centrifugal evaporator	Eppendorf (Hamburg, Germany)
Agarose gel chamber	PeqLab (Erlangen, Germany)
Vertical PAGE chamber (LSG-400-20)	C.B.S. Scientific (San Diego, USA)
Mini-PROTEAN Tetra Vertical Electrophoresis Cell	Bio-Rad (Feldkirchen, Germany)
Typhoon TRIO+ imager	GE Healthcare (Chicago, USA)
Tecan Infinite M200 PRO plate reader	Tecan (Männedorf, Switzerland)
Epson Perfection V700 Photo scanner	Epson (Suwa, Japan)
NanoDrop 2000 spectrophotometer	Thermo Fisher Scientific (Waltham, USA)
Agilent 2100 Bioanalyzer	Agilent Technologies (Santa Clara, USA)
Qubit 2.0 Fluorometer	Invitrogen (Carlsbad, USA)
MinION sequencing device	Oxford Nanopore Technologies (Oxford, UK)
PromethION P48 sequencing device	Oxford Nanopore Technologies (Oxford, UK)

5.1.2 Chemicals and Reagents

All chemicals and reagents used were of molecular biology grade or higher unless otherwise specified.

Chemical/Reagent	Supplier
Ammonium acetate (5 M)	Merck (Darmstadt, Germany)
Calcium chloride ($CaCl_2$)	Carl Roth (Karlsruhe, Germany)
Chloroform (HPLC grade)	Sigma-Aldrich (Steinheim, Germany)
Dimethyl sulfoxide (DMSO, anhydrous)	Sigma-Aldrich (Steinheim, Germany)
Ethanol (≥99.5%, absolute)	Carl Roth (Karlsruhe, Germany)
Formamide	Carl Roth (Karlsruhe, Germany)
Glycogen (5 mg/mL, RNA grade)	Thermo Fisher Scientific (Waltham, USA)
1-Methylimidazole	Thermo Fisher Scientific (Waltham, USA)
m^7GDP-imidazolide (m^7GDP-Im)	Hongene Biotech (Union City, USA)
Phenol (Roti-Aqua, pH 4.0, for RNA)	Carl Roth (Karlsruhe, Germany)
Sodium chloride (NaCl)	Carl Roth (Karlsruhe, Germany)
Agarose	Biozym (Hessisch Oldendorf, Germany)
Rotiphorese Gel 40% acrylamide (19:1)	Carl Roth (Karlsruhe, Germany)
TBE buffer (10×)	Carl Roth (Karlsruhe, Germany)
TEMED	Carl Roth (Karlsruhe, Germany)
GelRed nucleic acid stain (3×)	Biotium (Hayward, USA)
SYBR Gold nucleic acid stain (10,000×)	Thermo Fisher Scientific (Waltham, USA)
Coomassie Brilliant Blue R-250	Carl Roth (Karlsruhe, Germany)
PageRuler Prestained Protein Ladder	Thermo Fisher Scientific (Waltham, USA)
RiboRuler High Range RNA Ladder	Thermo Fisher Scientific (Waltham, USA)

5.1.3 Enzymes

Enzyme	Supplier
Eco53kI restriction enzyme	New England Biolabs (Ipswich, USA)
BamHI restriction enzyme (10 U/µL)	New England Biolabs (Ipswich, USA)
SwaI restriction enzyme	New England Biolabs (Ipswich, USA)
SacII restriction enzyme	New England Biolabs (Ipswich, USA)

DNase I (50 U/µL)	Thermo Fisher Scientific (Waltham, USA)
E. coli poly(A) polymerase	New England Biolabs (Ipswich, USA)
T4 DNA ligase (2 U/µL)	Thermo Fisher Scientific (Waltham, USA)
T4 RNA Ligase 2 (10 U/µL)	New England Biolabs (Ipswich, USA)
T4 polynucleotide kinase (10 U/µL)	Thermo Fisher Scientific (Waltham, USA)
T7 RNA polymerase (20 U/µL)	Thermo Fisher Scientific (Waltham, USA)
XRN-1 (5′→3′ exoribonuclease)	New England Biolabs (Ipswich, USA)

5.1.4 Kits

Kit	Supplier
HiScribe T7 High Yield RNA Synthesis Kit	New England Biolabs (Ipswich, USA)
Monarch RNA Cleanup Kit	New England Biolabs (Ipswich, USA)
Oligo Clean & Concentrator Kit	Zymo Research (Irvine, USA)
Qubit dsDNA HS Assay Kit	Invitrogen (Carlsbad, USA)
SQK-RNA002 Direct RNA Sequencing Kit	Oxford Nanopore Technologies (Oxford, UK)
SQK-RNA004 Direct RNA Sequencing Kit	Oxford Nanopore Technologies (Oxford, UK)
Flexi Rabbit Reticulocyte Lysate System	Promega (Walldorf, Germany)

5.1.5 Oligonucleotides

Chemically synthesized RNA oligonucleotides for splint ligation were obtained from the following sources. For m^5C-modified sequences used in the ModiCal training constructs, oligonucleotides were purchased from Dharmacon (Lafayette, USA). Unmodified RNA oligonucleotides and DNA splint oligonucleotides were purchased from Biomers (Ulm, Germany). For the SARS-CoV-2 and IRESeGFP5′ constructs, unmodified oligoribonucleotides were obtained from IBA (Göttingen, Germany), and 2′-O-methylated (Nm) oligonucleotides were purchased from Dharmacon (Lafayette, USA). Cy5-labeled oligonucleotides for chemical capping visualization were obtained from IBA (Göttingen, Germany). DNA splint oligonucleotides for the IRESeGFP5′ constructs were purchased from IDT (Coralville, USA) or IBA (Göttingen, Germany). All oligonucleotides were HPLC-purified and reconstituted in

nuclease-free water to a working concentration of 100 µM. Complete oligonucleotide sequences are provided in Appendix A.1.

5.1.6 Plasmid Vectors

Plasmid vectors for IVT were custom-synthesized and cloned into pUC57 backbones by GenScript (Piscataway, USA). Each plasmid contained the target RNA sequence downstream of a T7 promoter and carried an ampicillin resistance marker. Complete insert sequences are provided in Appendix A.3.

Plasmid Name	Vector	Restriction Enzyme	Description
pUC57_Sc25S	pUC57	BamHI	*S. cerevisiae* 25S rRNA (3396 nt)
IRESeGFP5′-S1_pUC57	pUC57	SacII	IRES-eGFP S1 mRNA, full length
IRESeGFP5′-1_pUC57	pUC57	BamHI	IRES-eGFP S1 mRNA, Fragment 1
IRESeGFP3′-1_pUC57	pUC57	SacII	IRES-eGFP S1 mRNA, Fragment 3
Nsp13-FL-pUC57	pUC57	Eco53kI	SARS-CoV-2 Nsp13/helicase full length
Nsp13-F1-pUC57	pUC57	Eco53kI	SARS-CoV-2 Nsp13/helicase Fragment 1
Nsp13-F3-pUC57	pUC57	Eco53kI	SARS-CoV-2 Nsp13/helicase Fragment 3
Spike-S1-FL-pUC57	pUC57	Eco53kI	SARS-CoV-2 Spike S1-NTD full length
Spike-S1-F1-pUC57	pUC57	SwaI	SARS-CoV-2 Spike S1-NTD Fragment 1
Spike-S1F3-pUC57	pUC57	Eco53kI	SARS-CoV-2 Spike S1-NTD Fragment 3

5.1.7 Yeast Strains

Saccharomyces cerevisiae strains used in this study included the wild-type BY4741 (MATa *his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0), the *rcm1*Δ deletion mutant (BY4741 rcm1::KanMX4), and the *nop2*Δ deletion mutant (BY4741 nop2::KanMX4). All strains were provided by Virginie Marchand and Yuri Motorin (Université de Lorraine, Nancy, France).

5.1.8 Buffers and Solutions

All buffers and solutions were prepared with ultrapure Milli-Q water unless otherwise stated.

Buffer/Solution	Composition
1× Kinase-ligase buffer	50 mM Tris-HCl (pH 7.4), 10 mM $MgCl_2$, 5 mM DTT, 2 mM ATP
PAGE loading buffer (denaturing)	1× TBE, 90% (v/v) formamide
8% denaturing PAGE (100 mL)	32 mL gel concentrate, 58 mL diluent, 10 mL buffer, 400 µL 10% APS, 50 µL TEMED
15% denaturing PAGE (100 mL)	60 mL gel concentrate, 30 mL diluent, 10 mL buffer, 400 µL 10% APS, 50 µL TEMED
20% denaturing PAGE (100 mL)	80 mL gel concentrate, 10 mL diluent, 10 mL buffer, 400 µL 10% APS, 50 µL TEMED
6× SDS loading buffer (non-reducing)	Prepared in-house; without β-mercaptoethanol
SDS running buffer (1×)	25 mM Tris, 192 mM glycine, 0.1% SDS

5.2 Methods

5.2.1 RNA Preparation

5.2.1.1 Yeast RNA Extraction

This step was performed by Virginie Marchand and Yuri Motorin (Université de Lorraine, Nancy, France). Total RNA was extracted from frozen yeast pellets using TRI Reagent according to the manufacturer's protocol. Cells were resuspended in 1 mL TRI Reagent per 10^7 cells, homogenized by vortexing, and processed through phase separation with chloroform. RNA was precipitated with isopropanol, washed with 75% ethanol, and resuspended in nuclease-free water. RNA

integrity was assessed by agarose gel electrophoresis and concentration was determined by NanoDrop spectrophotometry.

5.2.1.2 *In Vitro* Transcription (IVT)

Plasmid DNA was linearized overnight at 37 °C. Linearized DNA was purified by phenol-chloroform extraction. Twice with phenol:chloroform:isoamyl alcohol (25:24:1), twice with chloroform, followed by ethanol precipitation with 1/10 volume 5 M ammonium acetate, 1 µL glycogen (5 mg/mL), and 2.5 volumes ethanol. Precipitation was performed at –80 °C for 1 h or –20 °C overnight. DNA pellets were collected by centrifugation at 12,000 g and –4 °C for 45 min, washed with 75% cold ethanol, and centrifuged again for 15 min. The dried pellet was resuspended in ultrapure water. IVT was carried out using the HiScribe T7 High Yield RNA Synthesis Kit following the manufacturer's instructions with 4 h incubation at 37 °C. For transcripts requiring a 5′-monophosphate terminus, GTP concentration was reduced to 2 mM and GMP was added at 16 mM to initiate transcription with a monophosphate. DNA template was removed by DNase I digestion (2 U per µg DNA) at 37 °C for 30 min. RNA was purified using the Monarch RNA Cleanup Kit and assessed by agarose gel electrophoresis and NanoDrop spectrophotometry.

5.2.1.3 Three-Way Splint Ligation of Long Constructs

For the SARS-CoV-2 constructs (~960–1800 nt) and the IRESeGFP5′ test construct (1371 nt), long splint ligation was performed following the protocol established by Hertler, Slama et al. (2022). Two IVT-derived flanking fragments (Fragment #1 and Fragment #3) and a central 19-nucleotide synthetic oligoribonucleotide (Fragment #2, carrying either the modification of interest or an unmodified control) were assembled in a one-pot reaction. Equimolar amounts of the three RNA fragments were combined with a complementary DNA splint oligonucleotide in T4 RNA Ligase 2 reaction buffer. The mixture was heated to 75 °C for 4 min and cooled to room temperature over 15 min to allow annealing. T4 RNA Ligase 2 (10 U/µL) was added, and ligation was performed overnight at 16 °C. Residual DNA splint was removed by DNase I digestion (0.1 U/µL) at 37 °C for 30 min. Full-length ligation products were purified by real-time agarose gel elution. The ligation mixture was separated on a 1% agarose gel pre-stained with SYBR Gold (1:10,000), and the full-length band was excised under DarkReader blue light transillumination to avoid UV-induced RNA damage. RNA was recovered by ethanol precipitation and filtered through NanoSep 0.2 µm solid-phase filters. Complete oligonucleotide sequences for all constructs are listed in Appendix A.1.1.

5.2.1.4 Splint Ligation for Synthetic Training Constructs

Three-fragment splint ligation was used to generate site-specific m^5C-modified and unmodified RNA constructs (99 nt) for yeast C2278 and DENV C1218 training data. Equimolar RNA oligonucleotides (30 µM each: 5′ arm, central fragment with m^5C or C, 3′ arm) were phosphorylated with T4 polynucleotide kinase (0.75 U/µL) in 1× kinase-ligase buffer (50 mM Tris-HCl pH 7.4, 10 mM $MgCl_2$, 5 mM DTT, 2 mM ATP) for 1 h at 37 °C. Phosphorylated oligonucleotides were annealed to a slightly sub-stoichiometric amount (2% less) of complementary DNA splint by heating to 75 °C for 4 min followed by gradual cooling to room temperature over 15 min. Ligation was initiated by adding T4 DNA ligase (2 U/µL) and incubating overnight at 16 °C. Residual DNA splint was degraded with DNase I (0.1 U/µL) at 37 °C for 30 min. Ligated products were purified by electrophoresis on 8–10% denaturing polyacrylamide gels. RNA integrity and concentration were determined using a NanoDrop 2000 spectrophotometer. Oligonucleotide sequences are listed in Appendix A.1.1 and A.1.2.

5.2.1.5 Chemical Capping with m^7GDP-Imidazolide

Post-transcriptional 5′ cap installation was performed using m7GDP-imidazolide as a chemical cap donor, which uniquely accepts 5′-monophosphate RNA substrates (Abe et al., 2022). The 5′-phosphorylated oligoribonucleotide (100 pmol) was precipitated with ammonium acetate/ethanol and dried by SpeedVac centrifugation (4 min, 100 mbar, 30 °C). The dried pellet was resuspended directly in the capping mixture: m7GDP-Im (40 mM final, in anhydrous DMSO), 1-methylimidazole (2 M), and $CaCl_2$ (10 mM) in a total volume of 5 µL. The reaction was incubated at 55 °C for 3 h. Capped oligonucleotides were purified by 20% denaturing PAGE (20 × 20 cm gel, 10–20 W, 3 h). Bands were visualized using Cy5 scanning (633 nm excitation, 670 nm BP 30 emission filter) on the Typhoon TRIO+ imager.

5.2.2 *In Vitro* Translation

In vitro translation was performed using the Flexi Rabbit Reticulocyte Lysate (RRL) system. For each reaction, 1 pmol mRNA was heat-denatured at 65 °C for 3 min, placed on ice, and added to the RRL mixture according to the manufacturer's instructions. The final mRNA concentration was 20 nM in a total volume of 50 µL. Reactions were incubated at 30 °C for 90 min.

5.2.3 Fluorescence Detection

5.2.3.1 In-Gel Fluorescence

Translation products were analyzed by SDS-PAGE following the protocol established by Slama (2021). Aliquots (2 µL) were mixed with non-reducing SDS loading buffer (without β-mercaptoethanol) and loaded without pre-heating onto 10% SDS-polyacrylamide gels to preserve eGFP fluorescence. Electrophoresis was performed in the Mini-PROTEAN Tetra system at 65 V for 30 min followed by 120 V. Gels were scanned directly on the Typhoon TRIO+ imager using the blue laser (488 nm excitation, 526 nm SP emission filter) for eGFP detection. The PageRuler Prestained Protein Ladder was visualized on the same gel using the red laser (633 nm excitation). Quantification of eGFP signal intensity was performed using ImageJ, normalized to background.

5.2.3.2 Plate Reader Fluorescence

For quantitative fluorescence measurements, eGFP fluorescence intensity was recorded using a Tecan Infinite M200 PRO plate reader with excitation at 488 nm and emission at 520 nm.

5.2.4 RiboMethSeq

RiboMethSeq analysis was performed by Virginie Marchand and Yuri Motorin (Université de Lorraine, Nancy, France) following established Illumina-based protocols (Marchand et al., 2016). Total RNA or synthetic constructs were subjected to controlled alkaline hydrolysis, and the resulting fragment libraries were sequenced on an Illumina NextSeq 2000 platform. The ScoreC2 metric, calculated from fragment end count distributions in a ±6 nt window around each position, was used to quantify protection from cleavage, with values above 0.6 considered indicative of 2′-O-methylation.

5.2.5 Bisulfite Sequencing

This step was performed by Virginie Marchand and Yuri Motorin (Université de Lorraine, Nancy, France). Total RNA (~50 ng) from *S. cerevisiae* was subjected to bisulfite treatment according to published protocols. RNA was converted to a DNA library using the NEBNext Small RNA Library Prep Kit. Libraries were sequenced on an Illumina NextSeq 2000 with 2 × 50 nt paired-end reads. The proportion of residual non-deaminated cytosine (UtoC_MM score) was calculated to determine m^5C stoichiometry at each position.

5.2.6 Nanopore Direct RNA Sequencing (DRS)

5.2.6.1 Polyadenylation

RNA samples (synthetic constructs, IVT, and native) were 3′-polyadenylated using *E. coli* poly(A) polymerase following the manufacturer's instructions. Poly(A)-tailed RNA was purified with the Oligo Clean & Concentrator Kit and eluted in nuclease-free water.

5.2.6.2 Library Preparation and Sequencing

DRS libraries were prepared from 1 µg polyadenylated RNA using either the SQK-RNA002 kit (for early experiments) or the SQK-RNA004 kit (for the ModiCal workflow and all calibration data), following the manufacturer's protocol. Key optimizations included extending all incubation times by 5 min compared to the standard protocol, mixing beads by gentle pipetting rather than flicking, drying ethanol washes for exactly 5 min, and performing spotON loading by adding the library dropwise with capillary action at the port. Where indicated, yeast samples were multiplexed using SeqTagger barcodes and demultiplexed with the SeqTagger tool. Libraries were quantified using the Qubit dsDNA HS Assay Kit and loaded onto ONT MinION or PromethION R10.4.1 flow cells (FLO-MIN114 or FLO-PRO114M). Sequencing was performed for up to 72 h under MinKNOW software (v24.02.26). Raw signal data were recorded in POD5 format.

For DENV, full-length IVT genomic RNA and corresponding nanopore DRS data from native viral RNA were obtained from Wu et al. (2026).

5.2.7 ModiCal Calibration Workflow

RNA modification detection was performed using ModiDeC, a dual-input neural network (~20 M parameters) combining structured inception blocks and LSTM layers for RNA-modification classification. All data curation, training, and analysis steps were carried out using the epi2me-labs ModiDeC workflow (v5.3.1). ModiCal implements a three-stage calibration strategy as follows.

Stage 1: Baseline Curation and Training. POD5 files from modified and unmodified splint-ligated constructs were aligned with minimap2, producing BAM files. The ModiDeC Curation workflow extracted signal chunks centered on the target cytidine using POD5–BAM–FASTA inputs with settings. Flow cell type RNA004, training mode, modification mapping enabled, modification type m^5C, 512 chunks per .npz file. Read indices were sampled from 0 to 100,000. From the resulting files, 800 .npz files per class were randomly selected for training and 200 per class for

validation. Training was performed with batch size 128, k-mer model 9, and 4 epochs, yielding the baseline model. Baseline analysis on native RNA was performed using the ModiDeC Analysis workflow. Positions with m^5C frequency above 2% were defined as apparent false-positive (FP) sites.

Stage 2: Bulk False-Positive Calibration. Unmodified IVT RNA was curated at all FP positions identified in baseline analysis, yielding ≥600 IVT .npz files. These were added to the existing training pool (800 modified + 800 unmodified ligation .npz), maintaining the validation set unchanged. Retraining with identical hyperparameters yielded the BulkFP model.

Stage 3: Iterative Single-FP Calibration. Residual FPs were resolved individually. For each FP coordinate, unmodified IVT chunks were curated, yielding ~10 .npz files per site. Only 1 .npz per FP site was added to the training pool in each round (Cal1, Cal2, Cal3…). The model was retrained with unchanged hyperparameters. This process was repeated until no cytidine in the reference RNA exceeded the 2% FP threshold.

5.2.8 Bioinformatic Analysis

Basecalling was performed using Dorado basecaller. Reads were aligned to reference sequences (yeast 25S rRNA or DENV genomic RNA) using minimap2 with parameters optimized for DRS. BAM files were processed with samtools for sorting, filtering, and indexing (version details in Appendix A3.1). Modification frequency at each reference position was computed as the fraction of reads classified as modified by the ModiCal pipeline. Detailed bioinformatic commands, scripts, and analysis workflows are available at https://github.com/zoezrend/ModiCal_docs.

5.2.9 Dorado Modification Calling

For benchmarking, POD5 files were basecalled with Dorado v0.8.2 using the rna004_130bps_sup@v5.1.0 model with modification detection enabled (--modified-bases m5C). Basecalling was performed as follows:

```
dorado basecaller rna004_130bps_sup@v5.1.0 pod5/ --modified-bases m5C >
calls.bam
```

Aligned BAM files were processed with modkit v0.4.2 to generate per-site modification pileups. Modification calls were filtered at two probability thresholds:

```
modkit pileup aligned.bam m5C_050.bed --ref ref.fasta --modified-bases 5mC --
mod-threshold m:0.50
modkit pileup aligned.bam m5C_095.bed --ref ref.fasta --modified-bases 5mC --
mod-threshold m:0.95
```

The permissive threshold (0.50) was used for sensitivity analysis, while the stringent threshold (0.95) served as the primary filter for benchmarking against ModiCal predictions.

5.2.10 Computational Resources

All deep learning training was performed on a dedicated workstation; hardware and software specifications are detailed in Appendix A.3 (Tables 20 and 21). Training times ranged from 2–4 hours per model (4 epochs, ~1,600 training files).

5.2.11 Data and Code Availability

All sequencing data generated in this study have been deposited in the European Nucleotide Archive (ENA) under accession number PRJEB106093. Trained ModiDeC models are available via Zenodo: https://doi.org/10.5281/zenodo.18131677

ModiCal documentation, helper scripts, and step-by-step protocols are provided at https://github.com/zoezrend/ModiCal_docs.

The ModiDeC neural network workflows used for data curation, model training, and analysis are publicly available at

https://github.com/stegiopast/wf-modidec-data-curation

https://github.com/stegiopast/wf-modidec-analysis

https://github.com/stegiopast/wf-modidec-training

6. Bibliography

[1] Abe, N., Imaeda, A., Inagaki, M., Li, Z., Kawaguchi, D., Onda, K., Nakashima, Y., Uchida, S., Hashiya, F., Kimura, Y., & Abe, H. (2022). Complete chemical synthesis of minimal messenger RNA by efficient chemical capping reaction. ACS Chemical Biology, 17(6), 1308–1314. https://doi.org/10.1021/acschembio.1c00996

[2] Abebe, J. S., Price, A. M., Hayer, K. E., Mohr, I., & Weitzman, M. D. (2022). DRUMMER—rapid detection of RNA modifications through comparative nanopore sequencing. Bioinformatics, 38(11), 3113–3115. https://doi.org/10.1093/bioinformatics/btac274

[3] Acera Mateos, P., Sethi, A. J., Ravindran, A., Phetsouphanh, C., Balachandran, S., Kelleher, A. D., Stark, M., & Hutvagner, G. (2024). Prediction of m^6A and m^5C at single-molecule resolution reveals a transcriptome-wide co-occurrence of RNA modifications. Nature Communications, 15(1), 3899. https://doi.org/10.1038/s41467-024-47953-7

[4] Acera Mateos, P., Zhou, Y., Zarnack, K., & Eyras, E. (2023). Concepts and methods for transcriptome-wide prediction of chemical messenger RNA modifications with machine learning. Briefings in Bioinformatics, 24(3), bbad163. https://doi.org/10.1093/bib/bbad163

[5] Aguilo, F., Li, S., Balasubramaniyan, N., Sancho, A., Benko, S., Zhang, F., Vashisht, A., Rengasamy, M., Andino, B., Chen, C. H., Zhou, F., Qian, C., Zhou, M. M., Wohlschlegel, J. A., Zhang, W., Suchy, F. J., & Walsh, M. J. (2016). Deposition of 5-methylcytosine on enhancer RNAs enables the coactivator function of PGC-1α. Cell Reports, 14(3), 479–492. https://doi.org/10.1016/j.celrep.2015.12.043

[6] Alagna, N., Mündnich, S., Miedema, J., Helm, M., & Mertens, C. (2025). ModiDeC: A multi-RNA modification classifier for direct nanopore sequencing. Nucleic Acids Research, 53(14), gkaf673. https://doi.org/10.1093/nar/gkaf673

[7] Ayadi, L., Motorin, Y., & Marchand, V. (2018). Quantification of 2'-O-Me residues in RNA using next-generation sequencing (Illumina RiboMethSeq protocol). Methods in Molecular Biology, 1649, 29–48. https://doi.org/10.1007/978-1-4939-7213-5_2

[8] Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature, 226(5252), 1209–1211. https://doi.org/10.1038/2261209a0

[9] Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell, 116(2), 281–297. https://doi.org/10.1016/S0092-8674(04)00045-5

[10] Basanta-Sanchez, M., Temple, S., Ansari, S. A., D'Amico, A., & Agris, P. F. (2016). Attomole quantification and global profile of RNA modifications. Nucleic Acids Research, 44(1), e26. https://doi.org/10.1093/nar/gkv971

[11] Beverly, M., Dell, A., Parmar, P., & Houghton, L. (2016). Label-free analysis of mRNA capping efficiency using RNase H probes and LC-MS. Analytical and Bioanalytical Chemistry, 408(18), 5021–5030. https://doi.org/10.1007/s00216-016-9605-x

[12] Birkedal, U., Christensen-Dalsgaard, M., Krogh, N., Sabarinathan, R., Gorodkin, J., & Nielsen, H. (2015). Profiling of ribose methylations in RNA by high-throughput sequencing. Angewandte Chemie International Edition, 54(2), 451–455. https://doi.org/10.1002/anie.201408362

[13] Blanco, S., Dietmann, S., Flores, J. V., Hussain, S., Kutter, C., Humphreys, P., Lombard, P., Treps, L., Popis, M., Kellner, S., Hölter, S. M., Garrett, L., Wurst, W., Becker, L., Klopstock, T., Fuchs, H., Gailus-Durner, V., Hrabě de Angelis, M., Karadottir, R. T., Helm, M., Ule, J., Gleeson, J. G., Odom, D. T., & Frye, M. (2014). Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. The EMBO Journal, 33(18), 2020–2039. https://doi.org/10.15252/embj.201489843

[14] Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., Avsar, G., Romitelli, A., Pir, P., Dassi, E., Conticello, S. G., Aguilo, F., & Bujnicki, J. M. (2022). MODOMICS: A database of RNA modification pathways. 2021 update. Nucleic Acids Research, 50(D1), D231–D235. https://doi.org/10.1093/nar/gkab1083

[15] Bohnsack, K. E., Höbartner, C., & Bohnsack, M. T. (2019). Eukaryotic 5-methylcytosine (m^5C) RNA methyltransferases: Mechanisms, cellular functions, and links to disease. Genes, 10(2), 102. https://doi.org/10.3390/genes10020102

[16] Bourgeois, G., Ney, M., Gaspar, I., Aigueperse, C., Schaefer, M., Kellner, S., Helm, M., & Motorin, Y. (2015). Eukaryotic rRNA modification by yeast 5-methylcytosine-methyltransferases and human proliferation-associated antigen p120. PLOS ONE, 10(7), e0133321. https://doi.org/10.1371/journal.pone.0133321

[17] Boža, V., Brejová, B., & Vinař, T. (2017). DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS ONE, 12(6), e0178751. https://doi.org/10.1371/journal.pone.0178751

[18] Cámara, Y., Asin-Cayuela, J., Park, C. B., Metodiev, M. D., Shi, Y., Ruzzenente, B., Kukat, C., Habermann, B., Wibom, R., Hultenby, K., Franz, T., Erdjument-Bromage, H., Tempst, P., Hallberg, B. M., Gustafsson, C. M., & Larsson, N.-G. (2011). MTERF4 regulates translation by

targeting the methyltransferase NSUN4 to the mammalian mitochondrial ribosome. Cell Metabolism, 13(5), 527–539. https://doi.org/10.1016/j.cmet.2011.04.002

[19] Cech, T. R., Zaug, A. J., & Grabowski, P. J. (1981). In vitro splicing of the ribosomal RNA precursor of Tetrahymena. Cell, 27(3), 487–496. https://doi.org/10.1016/0092-8674(81)90390-1

[20] Chen, X., Li, A., Sun, B. F., Yang, Y., Han, Y. N., Yuan, X., Chen, R. X., Wei, W. S., Liu, Y., Gao, C. C., & Chen, Y. S. (2019). 5-Methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. Nature Cell Biology, 21(8), 978–990. https://doi.org/10.1038/s41556-019-0361-y

[21] Chen, Y. S., Yang, W. L., Zhao, Y. L., & Yang, Y. G. (2021). Dynamic transcriptomic m^{5}C and its regulatory role in RNA processing. Wiley Interdisciplinary Reviews: RNA, 12(4), e1639. https://doi.org/10.1002/wrna.1639

[22] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C., Alexandari, A. M., Lu, Z., Harris, D. J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Segler, M. H. S., Boca, S. M., Swamidass, S. J., Huang, A., Gitter, A., & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

[23] Choi, J., Indrisiunaite, G., DeMirci, H., Ieong, K.-W., Wang, J., Petrov, A., Prabhakar, A., Rechavi, G., Dominissini, D., He, C., Ehrenberg, M., & Puglisi, J. D. (2018). 2'-O-methylation in mRNA disrupts tRNA decoding during translation elongation. Nature Structural & Molecular Biology, 25(3), 208–216. https://doi.org/10.1038/s41594-018-0030-z

[24] Cohn, W. E., & Volkin, E. (1951). Nucleoside-5'-phosphates from ribonucleic acid. Nature, 167(4247), 483–484. https://doi.org/10.1038/167483a0

[25] Crain, P. F., & McCloskey, J. A. (1998). Applications of mass spectrometry to the characterization of oligonucleotides and nucleic acids. Current Opinion in Biotechnology, 9(1), 25–34. https://doi.org/10.1016/S0958-1669(98)80080-8

[26] Crick, F. H. C. (1958). On protein synthesis. Symposia of the Society for Experimental Biology, 12, 138–163. [No DOI — pre-DOI era. PMID: 13580867]

[27] Cruciani, S., & Novoa, E. M. (2025). The new era of single-molecule RNA modification detection through nanopore base-calling models. Nature Reviews Molecular Cell Biology, 27(1), 10–18. https://doi.org/10.1038/s41580-025-00896-3

[28] Cruciani, S., Delgado-Tejedor, A., Pryszcz, L. P., Medina, R., Llovera, L., & Novoa, E. M. (2025). De novo basecalling of RNA modifications at single molecule and nucleotide resolution. Genome Biology, 26(1), 38. https://doi.org/10.1186/s13059-025-03498-6

[29] Cui, X., Liang, Z., Shen, L., Zhang, Q., Bao, S., Geng, Y., Zhang, B., Leo, V., Vardy, L. A., Lu, T., Gu, X., & Yu, H. (2017). 5-Methylcytosine RNA methylation in Arabidopsis thaliana. Molecular Plant, 10(11), 1387–1399. https://doi.org/10.1016/j.molp.2017.09.013

[30] Daffis, S., Szretter, K. J., Schriewer, J., Li, J., Youn, S., Errett, J., Gangappa, S., Brassil, M. M., & Diamond, M. S. (2010). 2'-O methylation of the viral mRNA cap evades host restriction by IFIT family members. Nature, 468(7322), 452–456. https://doi.org/10.1038/nature09489

[31] Dai, Q., Moshitch-Moshkovitz, S., Han, D., Kol, N., Amariglio, N., Rechavi, G., Dominissini, D., & He, C. (2017). Nm-seq maps 2'-O-methylation sites in human mRNA with base precision. Nature Methods, 14(7), 695–698. https://doi.org/10.1038/nmeth.4294

[32] Dai, Q., Ye, C., Irkliyenko, I., Wang, Y., Sun, H. L., Gao, Y., He, C., & Sun, W. J. (2024). Ultrafast bisulfite sequencing detection of 5-methylcytosine in DNA and RNA. Nature Biotechnology, 42(10), 1559–1570. https://doi.org/10.1038/s41587-023-02034-w

[33] David, M., Dursi, L. J., Yao, D., Boutros, P. C., & Simpson, J. T. (2017). Nanocall: An open source basecaller for Oxford Nanopore sequencing data. Bioinformatics, 33(1), 49–55. https://doi.org/10.1093/bioinformatics/btw569

[34] Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. Nature Biotechnology, 34(5), 518–524. https://doi.org/10.1038/nbt.3423

[35] Decroly, E., Ferron, F., Lescar, J., & Canard, B. (2012). Conventional and unconventional mechanisms for capping viral mRNA. Nature Reviews Microbiology, 10(1), 51–65. https://doi.org/10.1038/nrmicro2675

[36] Delaunay, S., & Frye, M. (2019). RNA modifications regulating cell fate in cancer. Nature Cell Biology, 21(5), 552–559. https://doi.org/10.1038/s41556-019-0319-0

[37] Delaunay, S., Helm, M., & Frye, M. (2024). RNA modifications in physiology and disease: Towards clinical applications. Nature Reviews Genetics, 25(2), 104–122. https://doi.org/10.1038/s41576-023-00645-2

[38] Delazer, I., Bauer, I., Rummel, T., Schultze, M., Leitner, A., & Lusser, A. (2025). mRNA turnover dynamics are affected by cell differentiation and loss of the cytosine methyltransferase Nsun2. Nucleic Acids Research, 53(19), gkaf995. https://doi.org/10.1093/nar/gkaf995

[39] Derrington, I. M., Craig, J. M., Stava, E., Laszlo, A. H., Ross, B. C., Brinkerhoff, H., Nova, I. C., Doering, K., Tickman, B. I., Ronaghi, M., & Gundlach, J. H. (2015). Subangstrom single-molecule measurements of motor proteins using a nanopore. Nature Biotechnology, 33(10), 1073–1075. https://doi.org/10.1038/nbt.3357

[40] Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., & Sorek, R. (2012). Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. Nature, 485(7397), 201–206. https://doi.org/10.1038/nature11112

[41] Dong, H., Chang, D. C., Hua, M. H. C., Lim, S. P., Chionh, Y. H., Hia, F., Lee, Y. H., Kukkaro, P., Lok, S.-M., Dedon, P. C., & Shi, P.-Y. (2012). 2'-O methylation of internal adenosine by flavivirus NS5 methyltransferase. PLoS Pathogens, 8(4), e1002642. https://doi.org/10.1371/journal.ppat.1002642

[42] Dong, H., Gao, Y., Cai, Z., Li, Y., Li, X., Zhao, F., & Zhang, J. (2026). Comprehensive mapping of RNA modification dynamics and crosstalk via deep learning and nanopore direct RNA-sequencing. Nature Communications. https://doi.org/10.1038/s41467-026-68419-y

[43] Doudna, J. A., & Cech, T. R. (2002). The chemical repertoire of natural ribozymes. Nature, 418(6894), 222–228. https://doi.org/10.1038/418222a

[44] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[45] Fleming, A. M., & Burrows, C. J. (2026). Probing the epitranscriptome and RNA damage with nanopore direct RNA sequencing. RNA, advance online publication. https://doi.org/10.1261/rna.080908.125

[46] Fleming, A. M., Bommisetti, P., Xiao, S., Bandarian, V., & Burrows, C. J. (2023). Direct nanopore sequencing for the 17 RNA modification types in 36 locations in the E. coli ribosome enables monitoring of stress-dependent changes. ACS Chemical Biology, 18(10), 2211–2223. https://doi.org/10.1021/acschembio.3c00166

[47] Fleming, A. M., Mathewson, N. J., Howpay Manage, S. A., & Burrows, C. J. (2021). Nanopore dwell time analysis permits sequencing and conformational assignment of pseudouridine in SARS-CoV-2. ACS Central Science, 7(10), 1707–1717. https://doi.org/10.1021/acscentsci.1c00788

[48] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences, 89(5), 1827–1831. https://doi.org/10.1073/pnas.89.5.1827

[49] Frye, M., Harada, B. T., Behm, M., & He, C. (2018). RNA modifications modulate gene expression during development. Science, 361(6409), 1346–1349. https://doi.org/10.1126/science.aau1646

[50] Fu, L., Guerrero, C. R., Zhong, N., Amato, N. J., Liu, Y., Liu, S., Cai, Q., Ji, D., Jin, S. G., Niedernhofer, L. J., & Pfeifer, G. P. (2014). Tet-mediated formation of 5-hydroxymethylcytosine in RNA. Journal of the American Chemical Society, 136(33), 11582–11585. https://doi.org/10.1021/ja505305z

[51] Gamaarachchi, H., Ferguson, J. M., Samarakoon, H., Liyanage, K., & Deveson, I. W. (2024). Simulation of nanopore sequencing signal data with tunable parameters. Genome Research, 34(5), 778–783. https://doi.org/10.1101/gr.278730.123

[52] Gao, Y., Liu, X., Wu, B., Wang, H., Xi, F., Kohnen, M. V., Reddy, A. S., & Gu, L. (2021). Quantitative profiling of N^6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. Genome Biology, 22(1), 22. https://doi.org/10.1186/s13059-020-02241-7

[53] Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., Sherwood, R., & Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nature Methods, 15(3), 201–206. https://doi.org/10.1038/nmeth.4577

[54] Garcia-Campos, M. A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., Winkler, R., Nir, R., Lasman, L., Brandis, A., Hanna, J. H., Rossmanith, W., & Schwartz, S. (2019). Deciphering the "m^{6}A code" via antibody-independent quantitative profiling. Cell, 178(3), 731–747. https://doi.org/10.1016/j.cell.2019.06.013

[55] Gilbert, W. (1986). Origin of life: The RNA world. Nature, 319(6055), 618. https://doi.org/10.1038/319618a0

[56] Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C. L., Zhang, X., Golic, K. G., Jacobsen, S. E., & Bestor, T. H. (2006). Methylation of tRNA^Asp by the DNA methyltransferase homolog Dnmt2. Science, 311(5759), 395–398. https://doi.org/10.1126/science.1120976

[57] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[58] Grosjean, H. (2009). Nucleic acids are not boring long polymers of only four types of nucleotides: A guided tour. In H. Grosjean (Ed.), DNA and RNA modification enzymes: Structure, mechanism, function and evolution (pp. 1–18). Landes Bioscience.

[59] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., & Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell, 35(3), 849–857. https://doi.org/10.1016/0092-8674(83)90117-4

[60] Guo, Z., Shao, Y., Tan, L., Lu, B., Deng, X., Chen, S., & Li, R. (2025). Enhanced detection of RNA modifications in Escherichia coli utilizing direct RNA sequencing. Cell Reports Methods, 5(9), 101168. https://doi.org/10.1016/j.crmeth.2025.101168

[61] Guseva, E. A., Averina, O. A., Isaev, S. V., Pletnev, P. I., Bragina, E. E., Permyakov, O. A., Buev, V. S., Priymak, A. V., Emelianova, M. A., Pshanichnaya, L., Romanov, E. A., Novikova, S. E., Petriukov, K. S., Golovina, A. Y., Grigorieva, O. O., Manskikh, V. N., Korshunova, D. S., Silaeva, Y. Y., Deykin, A. V., Rubtsova, M. P., Zgoda, V. G., Mazur, A. M., Prokhortchouk, E. B., Dontsova, O. A., & Sergiev, P. V. (2025). Positioning of sperm tail longitudinal columns depends on NSUN7, an RNA-binding protein destabilizing elongated spermatid transcripts. RNA, 31(5), 709–723. https://doi.org/10.1261/rna.080320.124

[62] Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. Nature, 482(7385), 339–346. https://doi.org/10.1038/nature10887

[63] Haag, S., Warda, A. S., Kretschmer, J., Günnigmann, M. A., Höbartner, C., & Bohnsack, M. T. (2015). NSUN6 is a human RNA methyltransferase that catalyzes formation of m5C72 in specific tRNAs. RNA, 21(9), 1532–1543. https://doi.org/10.1261/rna.051524.115

[64] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[65] Harris, T., Marquez, B., Suarez, S., & Schimenti, J. (2007). Sperm motility defects and infertility in male mice with a mutation in Nsun7, a member of the Sun domain-containing family of putative RNA methyltransferases. Biology of Reproduction, 77(2), 376–382. https://doi.org/10.1095/biolreprod.106.058669

[66] Heissenberger, C., Liendl, L., Nagelreiter, F., Gonskikh, Y., Yang, G., Stelzer, E. M., Krammer, T. L., Micutkova, L., Vogt, S., Kreil, D. P., Sekot, G., Siena, E., Poser, I., Harreither, E., Linder, A., Ehret, V., Helbich, T. H., Grillari-Voglauer, R., Jansen-Dürr, P., Kos, M., Polacek, N., & Grillari, J. (2019). Loss of the ribosomal RNA methyltransferase NSUN5 impairs global protein synthesis and normal growth. Nucleic Acids Research, 47(22), 11807–11825. https://doi.org/10.1093/nar/gkz1043

[67] Helm, M., & Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. Nature Reviews Genetics, 18(5), 275–291. https://doi.org/10.1038/nrg.2016.169

[68] Hendra, C., Pratanwanich, P. N., Wan, Y. K., Goh, W. S. S., Thiery, A., & Göke, J. (2022). Detection of m^{6}A from direct RNA sequencing using a multiple instance learning framework. Nature Methods, 19(12), 1590–1598. https://doi.org/10.1038/s41592-022-01666-1

[69] Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. Nature Reviews Molecular Cell Biology, 19(5), 327–341. https://doi.org/10.1038/nrm.2017.130

[70] Hertler, J., Slama, K., Schober, B., Özrendeci, Z., Marchand, V., Motorin, Y., & Helm, M. (2022). Synthesis of point-modified mRNA. Nucleic Acids Research, 50(20), e115. https://doi.org/10.1093/nar/gkac719

[71] Höbartner, C., Bohnsack, K. E., & Bohnsack, M. T. (2024). How natural enzymes and synthetic ribozymes generate methylated nucleotides in RNA. Annual Review of Biochemistry, 93(1), 109–137. https://doi.org/10.1146/annurev-biochem-030222-112310

[72] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[73] Hoernes, T. P., Clementi, N., Faserl, K., Glasner, H., Breuker, K., Lindner, H., Huttenhofer, A., & Erlacher, M. D. (2016). Nucleotide modifications within bacterial messenger RNAs regulate their translation and are able to rewire the genetic code. Nucleic Acids Research, 44(2), 852–862. https://doi.org/10.1093/nar/gkv1182

[74] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons.

[75] Huang, S., Wylder, A. C., & Pan, T. (2024). Simultaneous nanopore profiling of mRNA m6A and pseudouridine reveals translation coordination. Nature Biotechnology, 42(12), 1831–1835. https://doi.org/10.1038/s41587-024-02135-0

[76] Huang, T., Chen, W., Liu, J., Gu, N., & Zhang, R. (2019). Genome-wide identification of mRNA 5-methylcytosine in mammals. Nature Structural & Molecular Biology, 26(5), 380–388. https://doi.org/10.1038/s41594-019-0218-x

[77] Hussain, S., Aleksic, J., Blanco, S., Dietmann, S., & Frye, M. (2013). Characterizing 5-methylcytosine in the mammalian epitranscriptome. Genome Biology, 14(11), 215. https://doi.org/10.1186/gb4143

[78] Ibrahim, F., Oppelt, J., Maragkakis, M., & Mourelatos, Z. (2021). TERA-Seq: true end-to-end sequencing of native RNA molecules for transcriptome characterization. Nucleic Acids Research, 49(20), e115. https://doi.org/10.1093/nar/gkab713

[79] Jenjaroenpun, P., Wongsurawat, T., Wadley, T. D., Wassenaar, T. M., Liu, J., Dai, Q., He, C., & Nookaew, I. (2021). Decoding the epitranscriptional landscape from native RNA sequences. Nucleic Acids Research, 49(2), e7. https://doi.org/10.1093/nar/gkaa620

[80] Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences, 93(24), 13770–13773. https://doi.org/10.1073/pnas.93.24.13770

[81] Kawai, G., Yamamoto, Y., Kamimura, T., Masegi, T., Sekine, M., Hata, T., Iimori, T., Watanabe, T., Miyazawa, T., & Yokoyama, S. (1992). Conformational rigidity of specific pyrimidine residues in tRNA arises from posttranscriptional modifications that enhance steric interaction between the base and the 2'-hydroxyl group. Biochemistry, 31(4), 1040–1046. https://doi.org/10.1021/bi00119a012

[82] Kellner, S., Burhenne, J., & Helm, M. (2010). Detection of RNA modifications. RNA Biology, 7(2), 237–247. https://doi.org/10.4161/rna.7.2.11468

[83] Kershaw, C. J., & O'Keefe, R. T. (2012). Splint ligation of RNA with T4 DNA ligase. Methods in Molecular Biology, 941, 257–269. https://doi.org/10.1007/978-1-62703-113-4_19

[84] Khan, M. A., Rafiq, M. A., Noor, A., Hussain, S., Flores, J. V., Rupp, V., Vincent, A. K., Malli, R., Ali, G., Khan, F. S., Ishak, G. E., Doherty, D., Weksberg, R., Ayub, M., Windpassinger, C., Ibrahim, S., Frye, M., Ansar, M., & Vincent, J. B. (2012). Mutation in NSUN2, which encodes an RNA methyltransferase, causes autosomal-recessive intellectual disability. American Journal of Human Genetics, 90(5), 856–863. https://doi.org/10.1016/j.ajhg.2012.03.023

[85] Khoddami, V., & Cairns, B. R. (2013). Identification of direct targets and modified bases of RNA cytosine methyltransferases. Nature Biotechnology, 31(5), 458–464. https://doi.org/10.1038/nbt.2566

[86] Khosronezhad, N., Colagar, A. H., & Jorsaraei, S. G. (2015). The Nsun7 (A11337)-deletion mutation causes reduction of its protein rate and is associated with sperm motility defect in infertile men. Journal of Assisted Reproduction and Genetics, 32(5), 807–815. https://doi.org/10.1007/s10815-015-0443-0

[87] Kowalak, J. A., Pomerantz, S. C., Crain, P. F., & McCloskey, J. A. (1993). A novel method for the determination of post-transcriptional modification in RNA by mass spectrometry. Nucleic Acids Research, 21(19), 4577–4585. https://doi.org/10.1093/nar/21.19.4577

[88] Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., & Cech, T. R. (1982). Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell, 31(1), 147–157. https://doi.org/10.1016/0092-8674(82)90414-7

[89] LeCun, Y., Bengio, Y., & others (1995). Convolutional networks for images, speech, and time series. In M. A. Arbib (Ed.), The handbook of brain theory and neural networks (pp. 255–258). MIT Press.

[90] Lee, J., Ahn, S., Kim, H.-S., An, J., & Sim, J. (2024). A robust model training strategy using hard negative mining in a weakly labeled dataset for lymphatic invasion in gastric cancer. The Journal of Pathology: Clinical Research, 10(1), e355. https://doi.org/10.1002/cjp2.355

[91] Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., Migliori, V., Toolan-Kerr, P., Sideri, T., Enright, A. J., Tzelepis, K., van Werven, F. J., Luscombe, N. M., Barbieri, I., Ule, J., Fitzgerald, T., Birney, E., Leonardi, T., & Kouzarides, T. (2021). RNA modifications detection by comparative Nanopore direct RNA sequencing. Nature Communications, 12(1), 7198. https://doi.org/10.1038/s41467-021-27393-3

[92] Legrand, C., Tuorto, F., Hartmann, M., Liebers, R., Jacob, D., Helm, M., & Lyko, F. (2017). Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. Genome Research, 27(9), 1589–1596. https://doi.org/10.1101/gr.210666.116

[93] Leontis, N. B., & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. RNA, 7(4), 499–512. https://doi.org/10.1017/S1355838201002515

[94] Li, J., Zhu, W.-Y., Wang, Z., Zhang, Y.-Y., Hua, M.-M., Liu, Q.-Y., Zhou, Y.-J., Wu, X.-Y., Li, H., Liu, M.-F., & Liu, R.-J. (2025). NSUN7 is a catalytically inactive RNA m^5C methyltransferase essential for sperm flagellum assembly. Nature Communications, 17(1), 532. https://doi.org/10.1038/s41467-025-67233-2

[95] Liao, H., Gaur, A., McConie, H., Shekar, A., Wang, K., Chang, J. T., Breton, G., & Denicourt, C. (2022). Human NOP2/NSUN1 regulates ribosome biogenesis through non-catalytic complex formation with box C/D snoRNPs. Nucleic Acids Research, 50(18), 10695–10716. https://doi.org/10.1093/nar/gkac817

[96] Limbach, P. A., Crain, P. F., & McCloskey, J. A. (1994). Summary: The modified nucleosides of RNA. Nucleic Acids Research, 22(12), 2183–2196. https://doi.org/10.1093/nar/22.12.2183

[97] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980–2988. https://doi.org/10.1109/ICCV.2017.324

[98] Ling, C., Lebeau, B., Keong, K. C., & Fullwood, M. (2026). Advanced deep learning strategies in nanopore RNA sequencing. RNA Biology, 23(1), 1–15. https://doi.org/10.1080/15476286.2026.2627968

[99] Liu, H., Begik, O., & Novoa, E. M. (2021). EpiNano: Detection of m^{6}A RNA modifications using Oxford Nanopore direct RNA sequencing. Methods in Molecular Biology, 2284, 31–52. https://doi.org/10.1007/978-1-0716-1307-8_3

[100] Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., Schwartz, S., Mattick, J. S., Smith, M. A., & Novoa, E. M. (2019). Accurate detection of m^{6}A RNA modifications in native RNA sequences. Nature Communications, 10(1), 4079. https://doi.org/10.1038/s41467-019-11713-9

[101] Long, T., Li, J., Li, H., Zhou, M., Zhou, X. L., Liu, R. J., & Wang, E. D. (2016). Sequence-specific and shape-selective RNA recognition by the human RNA 5-methylcytosine methyltransferase NSun6. Journal of Biological Chemistry, 291(46), 24293–24303. https://doi.org/10.1074/jbc.M116.742569

[102] Lorenz, D. A., Sathe, S., Einstein, J. M., & Yeo, G. W. (2020). Direct RNA sequencing enables m^{6}A detection in endogenous transcript isoforms at base resolution. RNA, 26(1), 19–28. https://doi.org/10.1261/rna.072785.119

[103] Lu, L., Zhang, X., Zhou, Y., Shi, Z., Xie, X., Zhang, X., Gao, L., Fu, A., Liu, C., He, B., Xiong, X., Yin, Y., Wang, Q., Yi, C., & Li, X. (2024). Base-resolution m5C profiling across the mammalian transcriptome by bisulfite-free enzyme-assisted chemical labeling approach. Molecular Cell, 84(15), 2984–3000.e8. https://doi.org/10.1016/j.molcel.2024.06.021

[104] Luo, T., Xu, M., Wang, M., Chen, F., & Shi, J. (2025). Systematic evaluation of computational tools for multitype RNA modification detection using nanopore direct RNA sequencing. Nature Methods, 23(2), 438–450. https://doi.org/10.1038/s41592-025-02974-y

[105] Makhamreh, A., Tavakoli, S., Fallahi, A., Kang, X., Gamper, H., Nabizadehmashhadtoroghi, M., Jain, M., Hou, Y. M., Rouhanifard, S. H., & Wanunu, M. (2024). Nanopore signal deviations from pseudouridine modifications in RNA are sequence-specific: Quantification requires dedicated synthetic controls. Scientific Reports, 14(1), 22457. https://doi.org/10.1038/s41598-024-72994-9

[106] Marchand, V., Blanloeil-Oillo, F., Helm, M., & Motorin, Y. (2016). Illumina-based RiboMethSeq approach for mapping of 2'-O-Me residues in RNA. Nucleic Acids Research, 44(16), e135. https://doi.org/10.1093/nar/gkw547

[107] Mattick, J. S. (2004). RNA regulation: A new genetics? Nature Reviews Genetics, 5(4), 316–323. https://doi.org/10.1038/nrg1321

[108] McIntyre, A. B. R., Gokhale, N. S., Cerchietti, L., Jaffrey, S. R., Horner, S. M., & Mason, C. E. (2020). Limits in the detection of m6A changes using MeRIP/m6A-seq. Scientific Reports, 10(1), 6590. https://doi.org/10.1038/s41598-020-63355-3

[109] Mei, L., Shen, C., Miao, R., Wang, J. Z., Cao, M. D., Zhang, Y. S., Shi, L. H., Zhao, G. H., Wang, M. H., Wu, L. S., & Wei, J. F. (2020). RNA methyltransferase NSUN2 promotes gastric cancer cell proliferation by repressing p57^Kip2 by an m^{5}C-dependent manner. Cell Death & Disease, 11(4), 270. https://doi.org/10.1038/s41419-020-2487-z

[110] Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., & Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell, 149(7), 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003

[111] Moore, M. J., & Sharp, P. A. (1992). Site-specific modification of pre-mRNA: The 2'-hydroxyl groups at the splice sites. Science, 256(5059), 992–997. https://doi.org/10.1126/science.1589782

[112] Moore, P. B., & Steitz, T. A. (2002). The involvement of RNA in ribosome function. Nature, 418(6894), 229–235. https://doi.org/10.1038/418229a

[113] Motorin, Y., & Helm, M. (2011). RNA nucleotide methylation. Wiley Interdisciplinary Reviews: RNA, 2(5), 611–631. https://doi.org/10.1002/wrna.79

[114] Motorin, Y., & Helm, M. (2024). General principles and limitations for detection of RNA modifications by sequencing. Accounts of Chemical Research, 57(3), 275–288. https://doi.org/10.1021/acs.accounts.3c00529

[115] Motorin, Y., & Marchand, V. (2018). Methods for RNA modification mapping using deep sequencing: Established and new emerging technologies. Genes, 9(12), 621. https://doi.org/10.3390/genes9120621

[116] Motorin, Y., Lyko, F., & Helm, M. (2010). 5-Methylcytosine in RNA: Detection, enzymatic formation and biological functions. Nucleic Acids Research, 38(5), 1415–1430. https://doi.org/10.1093/nar/gkp1117

[117] Naarmann-de Vries, I. S., Naarmann, B., & Dieterich, C. (2023). JACUSA2—a framework for identification and quantification of differential RNA modifications from duplex sequencing data. Bioinformatics, 39(6), btad363. https://doi.org/10.1093/bioinformatics/btad363

[118] Nakano, S., Suzuki, T., Kawarada, L., Iwata, H., Asano, K., & Suzuki, T. (2016). NSUN3 methylase initiates 5-formylcytidine biogenesis in human mitochondrial tRNA^Met. Nature Chemical Biology, 12(7), 546–551. https://doi.org/10.1038/nchembio.2099

[119] Nandakumar, J., Ho, C. K., Lima, C. D., & Shuman, S. (2004). RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2. Journal of Biological Chemistry, 279(30), 31337–31347. https://doi.org/10.1074/jbc.M402394200

[120] Ni, P., Xu, J., Zhong, Z., Luo, F., & Wang, J. (2024). RNA m6A detection using raw current signals and basecalling errors from Nanopore direct RNA sequencing reads. Bioinformatics, 40(6), btae375. https://doi.org/10.1093/bioinformatics/btae375

[121] Nishikura, K. (2016). A-to-I editing of coding and non-coding RNAs by ADARs. Nature Reviews Molecular Cell Biology, 17(2), 83–96. https://doi.org/10.1038/nrm.2015.4

[122] Ortiz-Barahona, V., Soler, M., Davalos, V., Garcia-Prieto, C. A., Janin, M., Setien, F., Fernandez-Rebollo, I., Bech-Serra, J. J., De La Torre, C., Guil, S., Villanueva, A., Zhang, P.-H., Yang, L., Guarnacci, M., Schumann, U., Preiss, T., Balaseviciute, U., Montal, R., Llovet, J. M., & Esteller, M. (2023). Epigenetic inactivation of the 5-methylcytosine RNA methyltransferase NSUN7 is associated with clinical outcome and therapeutic vulnerability in liver cancer. Molecular Cancer, 22(1), 83. https://doi.org/10.1186/s12943-023-01785-z

[123] Pandey, M., Fernandez, M., Gentile, F., Isayev, O., Tropsha, A., Stern, A. C., & Cherkasov, A. (2022). The transformational role of GPU computing and deep learning in drug discovery. Nature Machine Intelligence, 4(3), 211–221. https://doi.org/10.1038/s42256-022-00463-x

[124] Parker, M. T., Barton, G. J., & Simpson, G. G. (2021). Yanocomp: Robust prediction of m6A modifications in individual nanopore direct RNA reads. bioRxiv. https://doi.org/10.1101/2021.06.15.448494

[125] Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J. W., Barton, G. J., & Simpson, G. G. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m^{6}A modification. eLife, 9, e49658. https://doi.org/10.7554/eLife.49658

[126] Pichot, F., Marchand, V., Ayadi, L., Bourguignon-Igel, V., Helm, M., & Motorin, Y. (2020). Holistic optimization of bioinformatic analysis pipeline for detection and quantification of 2'-O-methylations in RNA by RiboMethSeq. Frontiers in Genetics, 11, 38. https://doi.org/10.3389/fgene.2020.00038

[127] Pichot, F., Marchand, V., Helm, M., & Motorin, Y. (2022). Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets. Methods, 203, 311–321. https://doi.org/10.1016/j.ymeth.2022.03.007

[128] Piechotta, M., Naarmann-de Vries, I. S., Wang, Q., Altmüller, J., & Dieterich, C. (2022). RNA modification mapping with JACUSA2. Genome Biology, 23, 115. https://doi.org/10.1186/s13059-022-02676-0

[129] Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W. Q., Wan, Y. K., Hendra, C., Prabhakar, P., Narang, V., Rossouw, D., Goh, J. J. L., & Göke, J. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nature Biotechnology, 39(11), 1394–1402. https://doi.org/10.1038/s41587-021-00949-w

[130] Rinn, J. L., & Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. Annual Review of Biochemistry, 81, 145–166. https://doi.org/10.1146/annurev-biochem-051410-092902

[131] Robinson, J. D., Chuang, C.-Y., Sra, S., & Jegelka, S. (2021). Contrastive learning with hard negative samples. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.04592

[132] Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA modifications in gene expression regulation. Cell, 169(7), 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045

[133] Rowley, H. A., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1), 23–38. https://doi.org/10.1109/34.655647

[134] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[135] Saletore, Y., Meyer, K., Korlach, J., Vilfan, I. D., Jaffrey, S., & Mason, C. E. (2012). The birth of the Epitranscriptome: Deciphering the function of RNA modifications. Genome Biology, 13(10), 175. https://doi.org/10.1186/gb-2012-13-10-175

[136] Salman, S., & Liu, X. (2019). Overfitting mechanism and avoidance in deep neural networks. arXiv preprint, arXiv:1901.06566. https://doi.org/10.48550/arXiv.1901.06566

[137] Samarakoon, H., Wan, Y. K., Parameswaran, S., Goke, J., Gamaarachchi, H., & Deveson, I. W. (2025). Leveraging basecaller's move table to generate a lightweight k-mer model for nanopore sequencing analysis. Bioinformatics, 41(4), btaf111. https://doi.org/10.1093/bioinformatics/btaf111

[138] Sawai, H., Wakai, H., & Nakamura-Ozaki, A. (1999). Synthesis and reactions of nucleoside 5'-diphosphate imidazolide. A nonenzymatic capping agent for 5'-monophosphorylated oligoribonucleotides in aqueous solution. Journal of Organic Chemistry, 64(16), 5836–5840. https://doi.org/10.1021/jo990286u

[139] Schaefer, M., Pollex, T., Hanna, K., & Lyko, F. (2009). RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Research, 37(2), e12. https://doi.org/10.1093/nar/gkn954

[140] Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meuburger, M., Helm, M., & Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes & Development, 24(15), 1590–1595. https://doi.org/10.1101/gad.586710

[141] Schosserer, M., Minois, N., Angerer, T. B., Tavernarakis, N., & Grillari, J. (2015). Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. Nature Communications, 6, 6158. https://doi.org/10.1038/ncomms7158

[142] Schumann, U., Shafik, A., & Preiss, T. (2020). METTL3 gains R/W access to the epitranscriptome. Molecular Cell, 78(6), 1066–1075. https://doi.org/10.1016/j.molcel.2020.05.026

[143] Selmi, T., Hussain, S., Dietmann, S., Heiß, M., Borland, K., Flad, S., Carter, J.-M., Dennison, R., Huang, Y.-L., Kellner, S., Börnelöv, S., & Frye, M. (2021). Sequence- and structure-specific cytosine-5 mRNA methylation by NSUN6. Nucleic Acids Research, 49(2), 1006–1022. https://doi.org/10.1093/nar/gkaa1193

[144] Sharma, S., Yang, J., Watzinger, P., Kötter, P., & Entian, K. D. (2013). Yeast Nop2 and Rcm1 methylate C2870 and C2278 of the 25S rRNA, respectively. Nucleic Acids Research, 41(19), 9062–9076. https://doi.org/10.1093/nar/gkt679

[145] Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 761–769. https://doi.org/10.1109/CVPR.2016.89

[146] Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nature Methods, 14(4), 407–410. https://doi.org/10.1038/nmeth.4184

[147] Slama, K. (2021). [Doctoral dissertation, Johannes Gutenberg-Universität Mainz].

[148] Sloan, K. E., Warda, A. S., Sharma, S., Entian, K.-D., Lafontaine, D. L. J., & Bohnsack, M. T. (2017). Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. RNA Biology, 14(9), 1138–1152. https://doi.org/10.1080/15476286.2016.1259781

[149] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[150] Song, Z., Huang, D., Song, B., Chen, K., Song, Y., Liu, G., Su, J., de Magalhaes, J. P., Rigden, D. J., & Meng, J. (2021). Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. Nature Communications, 12, 4011. https://doi.org/10.1038/s41467-021-24313-3

[151] Sordyl, D., Boileau, E., Bernat, A., Maiti, S., Mukherjee, S., Moafinejad, S. N., Amiri Farsani, M., Shavina, A., Cappannini, A., Agostini, G., Conticello, S. G., Stefaniak, F., Dieterich, C., Purta, E., & Bujnicki, J. M. (2026). MODOMICS: a database of RNA modifications and related information. 2025 update and 20th anniversary. Nucleic Acids Research, 54(D1), D219–D225. https://doi.org/10.1093/nar/gkaf1284

[152] Squires, J. E., Patel, H. R., Nousch, M., Sibbrit, I., Hilton, D. J., McLaughlin, K. A., & Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Research, 40(11), 5023–5033. https://doi.org/10.1093/nar/gks144

[153] Stevens, A. (2001). 5'-Exoribonuclease 1: Xrn1. Methods in Enzymology, 342, 251–259. https://doi.org/10.1016/S0076-6879(01)42549-3

[154] Stoiber, M., Quick, J., Egan, R., Lee, J. E., Celniker, S., Neely, R. K., Loman, N., Pennacchio, L. A., & Brown, J. (2017). De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. bioRxiv, 094672. https://doi.org/10.1101/094672

[155] Sung, K.-K., & Poggio, T. (1998). Example-based learning for view-based human face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(1), 39–51. https://doi.org/10.1109/34.655648

[156] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science, 240(4857), 1285–1293. https://doi.org/10.1126/science.3287615

[157] Tang, A. D., Felton, C., Hrabeta-Robinson, E., Volden, R., Vollmers, C., & Brooks, A. N. (2024). Detecting haplotype-specific transcript variation in long reads with FLAIR2. Genome Biology, 25(1), 173. https://doi.org/10.1186/s13059-024-03301-y

[158] Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nature Communications, 11, 1438. https://doi.org/10.1038/s41467-020-15171-6

[159] Tang, H., Kim, D. R., & Xie, X. (2018). Automated pulmonary nodule detection using 3D deep convolutional neural networks. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 523–526. https://doi.org/10.1109/isbi.2018.8363630

[160] Temin, H. M., & Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature, 226(5252), 1211–1213. https://doi.org/10.1038/2261211a0

[161] Teng, H., Stoiber, M., Bar-Joseph, Z., & Kingsford, C. (2024). Detecting m6A RNA modification from nanopore sequencing using a semi-supervised learning framework. Genome Research, 34(11), 1987–1999. https://doi.org/10.1101/gr.278960.124

[162] Thüring, K., Schmid, K., Keller, P., & Helm, M. (2016). Analysis of RNA modifications by liquid chromatography–tandem mass spectrometry. Methods, 107, 48–56. https://doi.org/10.1016/j.ymeth.2016.03.019

[163] Trixl, L., & Lusser, A. (2019). The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. Wiley Interdisciplinary Reviews: RNA, 10(1), e1510. https://doi.org/10.1002/wrna.1510

[164] Tuorto, F., Herbst, F., Alerasool, N., Bender, S., Popp, O., Federico, G., Reitter, S., Liebers, R., Stoecklin, G., Gröne, H.-J., Dittmar, G., Glimm, H., & Lyko, F. (2015). The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. The EMBO Journal, 34(18), 2350–2362. https://doi.org/10.15252/embj.201591382

[165] Tuorto, F., Liebers, R., Musch, T., Schaefer, M., Hofmann, S., Kellner, S., Frye, M., Helm, M., Stoecklin, G., & Lyko, F. (2012). RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. Nature Structural & Molecular Biology, 19(9), 900–905. https://doi.org/10.1038/nsmb.2357

[166] Van Haute, L., Dietmann, S., Kremer, L., Hussain, S., Pearce, S. F., Powell, C. A., Rorbach, J., Lantaff, R., Blanco, S., Sauer, S., Kotzaeridou, U., Hoffmann, G. F., Memari, Y., Kolb-Kokocinski, A., Durbin, R., Mayr, J. A., Frye, M., Prokisch, H., & Minczuk, M. (2016). Deficient methylation and formylation of mt-tRNA^{Met} wobble cytosine in a patient carrying mutations in NSUN3. Nature Communications, 7, 12039. https://doi.org/10.1038/ncomms12039

[167] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008. https://arxiv.org/abs/1706.03762

[168] Viswanathan, T., Arya, S., Chan, S.-H., Qi, S., Dai, N., Misra, A., Park, J.-G., Oladunni, F., Kovalskyy, D., Hromas, R. A., Martinez-Sobrido, L., & Gupta, Y. K. (2020). Structural basis of RNA cap modification by SARS-CoV-2. Nature Communications, 11(1), 3718. https://doi.org/10.1038/s41467-020-17496-8

[169] Watson, K. J., Bromley, R. E., & Dunning Hotopp, J. C. (2025). Duplexed direct RNA sequencing protocol using polyadenylation and polyuridylation. Microbiology Resource Announcements, 14(3), e01041-24. https://doi.org/10.1128/mra.01041-24

[170] Will, C. L., & Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harbor Perspectives in Biology, 3(7), a003707. https://doi.org/10.1101/cshperspect.a003707

[171] Wu, C. C., Naarmann-de Vries, I. S., Hartmann, J., Nidoieva, Z., Kopietz, K., Marchand, V., Özrendeci, Z., Lindner, D., Schelchshorn, S., Flad, S., Frye, M., Papavasiliou, F. N., Schirmeister, T., Stoecklin, G., Schott, J., Motorin, Y., Tuorto, F., Dieterich, C., Helm, M., & Ruggieri, A. (2026). The two-step purification method ViREn identifies a single NSUN6-mediated 5-methylcytosine modification promoting dengue virus RNA genome turnover. Nucleic Acids Research, 54(2), gkag003. https://doi.org/10.1093/nar/gkag003

[172] Wu, Y., Shao, W., Yan, M., Wang, Y., Xu, P., Huang, G., Li, X., Gregory, B. D., Yang, J., Wang, H., & Yu, X. (2024). Transfer learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing. Nature Communications, 15, 4049. https://doi.org/10.1038/s41467-024-48437-4

[173] Xia, J., Wu, L., & Li, S. Z. (2022). ProGCL: Rethinking hard negative mining in graph contrastive learning. Proceedings of the 39th International Conference on Machine Learning (ICML 2022), PMLR 162, 24332–24346.

[174] Xie, S., Ding, L., Yu, Y., Liu, L., Sun, Y., Li, J., Ruan, J., & Zhu, Z. (2026). A dual context-aware basecaller for nanopore direct RNA sequencing. Nature Communications, 17(1), 680. https://doi.org/10.1038/s41467-026-68566-2

[175] Yang, X., Yang, Y., Sun, B. F., Chen, Y. S., Xu, J. W., Lai, W. Y., Li, A., Wang, X., Bhattarai, D. P., Xiao, W., & Sun, H. Y. (2017). 5-Methylcytosine promotes mRNA export—NSUN2 as the methyltransferase and ALYREF as an m^5C reader. Cell Research, 27(5), 606–625. https://doi.org/10.1038/cr.2017.55

[176] Yu, B., Nagae, G., Midorikawa, Y., Tatsuno, K., Dasgupta, B., Aburatani, H., & Ueda, H. (2024). m6ATM: A deep learning framework for demystifying the m6A epitranscriptome with Nanopore long-read RNA-seq data. Briefings in Bioinformatics, 25(6), bbae529. https://doi.org/10.1093/bib/bbae529

[177] Zhang, Y., Wu, Y., Ma, J., Wu, Y., Li, L., Wang, H., Jia, G., Rigden, D. J., Meng, J., Huang, D., & Chen, K. (2025). DirectRM: integrated detection of landscape and crosstalk between multiple RNA modifications using direct RNA sequencing. Nature Communications, 16, 9450. https://doi.org/10.1038/s41467-025-64495-8

[178] Zhu, Y., Pirnie, S. P., & Carmichael, G. G. (2017). High-throughput and site-specific identification of 2'-O-methylation sites using ribose oxidation sequencing (RibOxi-seq). RNA, 23(8), 1303–1314. https://doi.org/10.1261/rna.061549.117

[179] Zou, Y., He, L., Wang, S., & Chen, K. (2025). Benchmarking nanopore direct RNA sequencing for RNA modification detection. Briefings in Bioinformatics, 26(4), bbaf404. https://doi.org/10.1093/bib/bbaf404

List of Publications and Presentations

1. Manuscripts Submitted for Publication

• **Özrendeci, Z.**, Mündnich, S., Pastore, S., Wu, C. C., Marchand, V., Motorin, Y., Ruggieri, A., Gerber, S., & Helm, M. (2026). ModiCal: A targeted calibration workflow for site-specific m^5C validation by nanopore direct RNA sequencing. *Manuscript submitted for publication.*

2. Peer-Reviewed Journal Articles

• Wu, C. C., Naarmann-de Vries, I. S., Hartmann, J., Nidoieva, Z., Kopietz, K., Marchand, V., **Özrendeci, Z.**, Lindner, D., Schelchshorn, S., Flad, S., Frye, M., Papavasiliou, F. N., Schirmeister, T., Stoecklin, G., Schott, J., Motorin, Y., Tuorto, F., Dieterich, C., Helm, M., & Ruggieri, A. (2026). The two-step purification method ViREn identifies a single NSUN6-mediated 5-methylcytosine modification promoting dengue virus RNA genome turnover. *Nucleic Acids Research, 54*(2), gkag003. https://doi.org/10.1093/nar/gkag003

• Hertler, J., Slama, K., Schober, B., **Özrendeci, Z.**, Marchand, V., Motorin, Y., & Helm, M. (2022). Synthesis of point-modified mRNA. *Nucleic Acids Research, 50*(20), e115. https://doi.org/10.1093/nar/gkac719

3. Conference Presentations

• *4th Symposium on Nucleic Acid Modifications*, Mainz, Germany, September 2024 (Poster)

• *12th Meeting of the GBM Study Section 'RNA-Biochemistry'*, Bonn, Germany, October 2023 (Flash Talk + Poster)

• *11th RMU-RNA Salon*, Gießen, Germany, September 2023 (Poster)

• *RNA and DNA Editing and Epitranscriptomics Across Biological Systems*, Ventura, CA, USA, March 2023 (Poster)

• *Vienna RNA Conference: RNA Modification and Processing*, Vienna, Austria, October 2022 (Poster)

Appendix

A.1 Oligonucleotide Sequences

A.1.1 Long Ligation Constructs

These constructs were used in long splint ligations for RiboMethSeq Nm validation and Nm translational studies.

A.1.1.1 Nsp13/Helicase Construct (1800 nt)

Component	Length	Sequence (5'→3')
F2 (unmodified)	23 nt	AAUUACAUCUUUCAUGGGAAGUU
F2 (Um501)	24 nt	AAUUACAUmCUUUCAUGGGAAGUU
cDNA splint	80 nt	TTTCGGTTAAGTGGTGGTCTAGGTTTACCAACTTCCCATGAAAGATGTAATTCTCTG TCAGACAGCACTTCACGTACAGT

A.1.1.2 Spike S1-NTD Construct (960 nt)

Component	Length	Sequence (5'→3')
F2 (unmodified)	23 nt	CAAUUUUGUAAUGAUCCAUUUUU
F2 (Am383)	23 nt	CAAUUUUGUAAmUGAUCCAUUUUU
cDNA splint	80 nt	ATTACAACAATAATTTCAGACACTTAAAGTTAAAACATTACTAGGTAAAAACCCACA AATAATGGTGTTTTTGTTGTTTT

A.1.1.3 IRESeGFP Construct (1371 nt)

Name	Modification	Sequence (5'→3')	Supplier
MH 905	unmodified	CCACAACCAUGGUGAGCAA	IBA
MH 1052	Nm (pos. 9)	CCACAACCAmUGGUGAGCAA	IBA
MH 1053	Nm (pos. 10)	CCACAACCAUmGGUGAGCAA	IBA
MH 1054	Nm (pos. 11)	CCACAACCAUGmGUGAGCAA	IBA
MH 1055	Nm (pos. 12)	CCACAACCAUGGmUGAGCAA	IBA
MH 1056	Nm (pos. 13)	CCACAACCAUGGUmGAGCAA	IBA
MH 1057	Nm (pos. 14)	CCACAACCAUGGUGmAGCAA	IBA
MH 1058	Nm (pos. 15)	CCACAACCAUGGUGAmGCAA	IBA
MH 1059	Nm (pos. 16)	CCACAACCAUGGUGAGmCAA	IBA
MH 1060	Nm (pos. 17)	CCACAACCAUGGUGAGCmAA	IBA
MH 1099	Nm (pos. 7)	CCACAAmCCAUGGUGAGCAA	IBA

MH 1229	dT-Cy5 (pos. 10)	CCACAACCA(dT-Cy5)GGUGAGCAA	Biomers
MH 1168	Splint cDNA	TGGGCACCACCCCGGTGAACAGCTCCTCGCCCTTGCTCACCATGGTT GTGGGATCCAAGCTTATCATCGTGTTTTTCAAAGG	IDT

A.1.2 Short Ligation Constructs

These constructs were used in short splint ligations to generate ground truth training data for ModiCal.

A.1.2.1 Yeast 25S rRNA C2278 Training Constructs (99 nt)

Oligonucleotide	Sequence (5′→3′)
Left arm (5′ arm)	ACCAAGCGCGGGUAAACGGCGGGAGUAACUAUGACUCUCU
Middle (m^5C-modified)	UAAGGUAGCm^5CAAAUGCCUC
Middle (unmodified)	UAAGGUAGCCAAAUGCCUC
Right arm (3′ arm)	GUCAUCUAAUUAGUGACGCGCAUGAAUGGAUUAACGAGAU
DNA splint	ATCCATTCATGCGCGTCACTAATTAGATGACGAGGCATTTGGCTACCTTAAGAGAGTCATAGTTACTC CCGCCGTTTACCCG

A.1.2.2 DENV gRNA C1218 Training Constructs (99 nt)

Oligonucleotide	Sequence (5′→3′)
Left arm (5′ arm)	GAACCCAGCCUAAAUGAAGAGCAGGACAAAAGGUUCGUCU
Middle (m^5C-modified)	GCAAACAm^5CUCCAUGGUGGA
Middle (unmodified)	GCAAACACUCCAUGGUGGA
Right arm (3′ arm)	CAGAGGAUGGGGAAAUGGAUGUGGAUUAUUUGGAAAAGGA
DNA splint	AAATAATCCACATCCATTTCCCCATCCTCTGTCCACCATGGAGTGTTTGCAGACGAACCTTTTGTCCT GCTCTTCATTTAGG

A.2 Plasmid Insert Sequences

The sequence is shown in the 5′→3′ direction (RNA format). Lowercase letters indicate the T7 promoter (5′ end) and poly(A) tail (3′ end).

25S rRNA insert (RNA, 5′→3′):

uaauacgacucacuauagggUUUGACCUCAAAUCAGGUAGGAGUACCCGCUGAACUUAAGCAUAUCAAUAAGCGGAGGAAAAGAAACCAACCGG
GAUUGCCUUAGUAACGGCGAGUGAAGCGGCAAAAGCUCAAAUUUGAAAUCUGGUACCUUCGGUGCCCGAGUUGUAAUUUGGAGAGGGCAACUUU
GGGGCCGUUCCUUGUCUAUGUUCCUUGGAACAGGACGUCAUAGAGGGUGAGAAUCCUGUGGCGAGGAGUGCGGUUCUUUGUAAAGUGCCUUCGA
AGAGUCGAGUUGUGUUGGGAAUGCAGCUCUAAGUGGGUGGUAAAUUCCAUCUAAAGCUAAAUAUUGGCGAGAGACCGAUAGCGAACAAGUACAG
UGAUGGAAAGAUGAAAAGAACUUUGAAAAGAGAGUGAAAAAGUACGUGAAAUUGUUGAAAGGGAAGGGCAUUUGAUCAGACAUGGUGUUUUGUG
CCCUCUGCUCCUUGUGGGUAGGGGAAUCUCGCAUUUCACUGGGCCAGCAUCAGUUUUGGUGGCAGGAUAAAUCCAUAGGAAUGUAGCUUGCCUC

GGUAAGUAUUAUAGCCUGUGGGAAUACUGCCAGCUGGGACUGAGGACUGCGACGUAAGUCAAGGAUGCUGGCAUAAUGGUUAUAUGCCGCCCGU
CUUGAAACACGGACCAAGGAGUCUAACGUCUAUGCGAGUGUUUGGGUGUAAAACCCAUACGCGUAAUGAAAGUGAACGUAGGUUGGGGCCUCGC
AAGAGGUGCACAAUCGACCGAUCCUGAUGUCUUCGGAUGGAUUUGAGUAAGAGCAUAGCUGUUGGGACCCGAAAGAUGGUGAACUAUGCCUGAA
UAGGGUGAAGCCAGAGGAAACUCUGGUGGAGGCUCGUAGCGGUUCUGACGUGCAAAUCGAUCGUCGAAUUUGGGUAUAGGGGCGAAAGACUAAU
CGAACCAUCUAGUAGCUGGUUCCUGCCGAAGUUUCCCUCAGGAUAGCAGAAGCUCGUAUCAGUUUUAUGAGGUAAAGCGAAUGAUUAGAGGUUC
CGGGGUGGAAAUGACCUUGACCUAUUCUCAAACUUUAAAUAUGUAAGAAGUCCUUGUUACUUAAUUGAACGUGGACAUUUGAAUGAAGAGCUUU
UAGUGGGCCAUUUUUGGUAAGCAGAACUGGCGAUGCGGGAUGAACCGAACGUAGAGUUAAGGUGCCGGAAUACACGCUCAUCAGACACCACAAA
AGGUGUUAGUUCAUCUAGACAGCCGGACGGUGGGCAUGGAAGUCGGAAUCCGCUAAGGAGUGUGUAACAACUCACCGGCCGAAUGAACUAGCCC
UGAAAAUGGAUGGCGCUCAAGCGUGUUACCUAUACUCUACCGUCAGGGUUGAUAUGAUGCCCUGACGAGUAGGCAGGCGUGGAGGUCAGUGACG
AAGCCUAGACCGUAAGGUCGGGUCGAACGGCCUCUAGUGCAGAUCUUGGUGGUAGUAGCAAAUAUUCAAAUGAGAACUUUGAAGACUGAAGUGG
GGAAAGGUUCCACGUCAACAGCAGUUGGACGUGGGUUAGUCGAUCCUAAGAGAUGGGGAAGCUCCGUUUCAAAGGCCUGAUUUUAUGCAGGCCA
CCAUCGAAAGGGAAUCCGGUUAAGAUUCCGGAACCUGGAUAUGGAUUCUUCACGGUAACGUAACUGAAUGUGGAGACGUCGGCGCGAGCCCUGG
GAGGAGUUAUCUUUUCUUCUUAACAGCUUAUCACCCCGGAAUUGGUUUAUCSGGAGaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaaaaaaa

IRES-eGFP S1 mRNA, full length (1370 nt):

GGGCGAAUUGGGUACCGGGCCCCCCCUCGAGGUCAUCGAAUUCCGCCCCUCUCCCUCCCCCCCCCCUAACGUUACUGGCCGAAGCCGCUUGGAA
UAAGGCCGGUGUGCGUUUGUCUAUAUGUUAUUUUCCACCAUAUUGCCGUCUUUUGGCAAUGUGAGGGCCCGGAAACCUGGCCCUGUCUUCUUGA
CGAGCAUUCCUAGGGGUCUUUCCCCUCUCGCCAAAGGAAUGCAAGGUCUGUUGAAUGUCGUGAAGGAAGCAGUUCCUCUGGAAGCUUCUUGAAG
ACAAACAACGUCUGUAGCGACCCUUUGCAGGCAGCGGAACCCCCCACCUGGCGACAGGUGCCUCUGCGGCCAAAAGCCACGUGUAUAAGAUACA
CCUGCAAAGGCGGCACAACCCCAGUGCCACGUUGUGAGUUGGAUAGUUGUGGAAAGAGUCAAAUGGCUCUCCUCAAGCGUAUUCAACAAGGGGC
UGAAGGAUGCCCAGAAGGUACCCCAUUGUAUGGGAUCUGAUCUGGGGCCUCGGUGCACAUGCUUUACAUGUGUUUAGUCGAGGUUAAAAAACGU
CUAGGCCCCCCGAACCACGGGGACGUGGUUUUCCUUUGAAAAACACGAUGAUAAGCUUGGAUCCCACAACCAUGGUGAGCAAGGGCGAGGAGCU
GUUCACCGGGGUGGUGCCCAUCCUGGUCGAGCUGGACGGCGACGUAAACGGCCACAAGUUCAGCGUGUCCGGCGAGGGCGAGGGCGAUGCCACC
UACGGCAAGCUGACCCUGAAGUUCAUCUGCACCACCGGCAAGCUGCCCGUGCCCUGGCCCACCCUCGUGACCACCCUGACCUACGGCGUGCAGU
GCUUCAGCCGCUACCCCGACCACAUGAAGCAGCACGACUUCUUCAAGUCCGCCAUGCCCGAAGGCUACGUCCAGGAGCGCACCAUCUUCUUCAA
GGACGACGGCAACUACAAGACCCGCGCCGAGGUGAAGUUCGAGGGCGACACCCUGGUGAACCGCAUCGAGCUGAAGGGCAUCGACUUCAAGGAG
GACGGCAACAUCCUGGGGCACAAGCUGGAGUACAACUACAACAGCCACAACGUCUAUAUCAUGGCCGACAAGCAGAAGAACGGCAUCAAGGUGA
ACUUCAAGAUCCGCCACAACAUCGAGGACGGCAGCGUGCAGCUCGCCGACCACUACCAGCAGAACACCCCCAUCGGCGACGGCCCCGUGCUGCU
GCCCGACAACCACUACCUGAGCACCCAGUCCGCCCUGAGCAAAGACCCCAACGAGAAGCGCGAUCACAUGGUCCUGCUGGAGUUCGUGACCGCC
GCCGGGAUCACUCUCGGCAUGGACGAGCUGUACAAGUAAAGCGGCCGCCACCGC

IRES-eGFP S1 mRNA, fragment 1 (627 nt):

GGGCGAAUUGGGUACCGGGCCCCCCCUCGAGGUCAUCGAAUUCCGCCCCUCUCCCUCCCCCCCCCCUAACGUUACUGGCCGAAGCCGCUUGGAA
UAAGGCCGGUGUGCGUUUGUCUAUAUGUUAUUUUCCACCAUAUUGCCGUCUUUUGGCAAUGUGAGGGCCCGGAAACCUGGCCCUGUCUUCUUGA
CGAGCAUUCCUAGGGGUCUUUCCCCUCUCGCCAAAGGAAUGCAAGGUCUGUUGAAUGUCGUGAAGGAAGCAGUUCCUCUGGAAGCUUCUUGAAG
ACAAACAACGUCUGUAGCGACCCUUUGCAGGCAGCGGAACCCCCCACCUGGCGACAGGUGCCUCUGCGGCCAAAAGCCACGUGUAUAAGAUACA
CCUGCAAAGGCGGCACAACCCCAGUGCCACGUUGUGAGUUGGAUAGUUGUGGAAAGAGUCAAAUGGCUCUCCUCAAGCGUAUUCAACAAGGGGC
UGAAGGAUGCCCAGAAGGUACCCCAUUGUAUGGGAUCUGAUCUGGGGCCUCGGUGCACAUGCUUUACAUGUGUUUAGUCGAGGUUAAAAAACGU
CUAGGCCCCCCGAACCACGGGGACGUGGUUUUCCUUUGAAAAACACGAUGAUAAGCUUGGAUC

IRES-eGFP S1 mRNA, fragment 3 (724 nt):

GGGCGAGGAGCUGUUCACCGGGGUGGUGCCCAUCCUGGUCGAGCUGGACGGCGACGUAAACGGCCACAAGUUCAGCGUGUCCGGCGAGGGCGAG
GGCGAUGCCACCUACGGCAAGCUGACCCUGAAGUUCAUCUGCACCACCGGCAAGCUGCCCGUGCCCUGGCCCACCCUCGUGACCACCCUGACCU
ACGGCGUGCAGUGCUUCAGCCGCUACCCCGACCACAUGAAGCAGCACGACUUCUUCAAGUCCGCCAUGCCCGAAGGCUACGUCCAGGAGCGCAC
CAUCUUCUUCAAGGACGACGGCAACUACAAGACCCGCGCCGAGGUGAAGUUCGAGGGCGACACCCUGGUGAACCGCAUCGAGCUGAAGGGCAUC
GACUUCAAGGAGGACGGCAACAUCCUGGGGCACAAGCUGGAGUACAACUACAACAGCCACAACGUCUAUAUCAUGGCCGACAAGCAGAAGAACG
GCAUCAAGGUGAACUUCAAGAUCCGCCACAACAUCGAGGACGGCAGCGUGCAGCUCGCCGACCACUACCAGCAGAACACCCCCAUCGGCGACGG
CCCCGUGCUGCUGCCCGACAACCACUACCUGAGCACCCAGUCCGCCCUGAGCAAAGACCCCAACGAGAAGCGCGAUCACAUGGUCCUGCUGGAG
UUCGUGACCGCCGCCGGGAUCACUCUCGGCAUGGACGAGCUGUACAAGUAAAGCGGCCGCCACCGC

Nsp13/helicase full-length construct (1803 nt):

GCUGUUGGGGCUUGUGUUCUUUGCAAUUCACAGACUUCAUUAAGAUGUGGUGCUUGCAUACGUAGACCAUUCUUAUGUUGUAAAUGCUGUUACG
ACCAUGUCAUAUCAACAUCACAUAAAUUAGUCUUGUCUGUUAAUCCGUAUGUUUGCAAUGCUCCAGGUUGUGAUGUCACAGAUGUGACUCAACU
UUACUUAGGAGGUAUGAGCUAUUAUUGUAAAUCACAUAAACCACCCAUUAGUUUUCCAUUGUGUGCUAAUGGACAAGUUUUUGGUUUAUAUAAA
AAUACAUGUGUUGGUAGCGAUAAUGUUACUGACUUUAAUGCAAUUGCAACAUGUGACUGGACAAAUGCUGGUGAUUACAUUUUAGCUAACACCU
GUACUGAAAGACUCAAGCUUUUUGCAGCAGAAACGCUCAAAGCUACUGAGGAGACAUUUAAACUGUCUUAUGGUAUUGCUACUGUACGUGAAGU
GCUGUCUGACAGAGAAUUACAUCUUUCAUGGGAAGUUGGUAAACCUAGACCACCACUUAACCGAAAUUAUGUCUUUACUGGUUAUCGUGUAACU
AAAAACAGUAAAGUACAAAUAGGAGAGUACACCUUUGAAAAAGGUGACUAUGGUGAUGCUGUUGUUUACCGAGGUACAACAACUUACAAAUUAA
AUGUUGGUGAUUAUUUUGUGCUGACAUCACAUACAGUAAUGCCAUUAAGUGCACCUACACUAGUGCCACAAGAGCACUAUGUUAGAAUUACUGG
CUUAUACCCAACACUCAAUAUCUCAGAUGAGUUUUCUAGCAAUGUUGCAAAUUAUCAAAAGGUUGGUAUGCAAAAGUAUUCUACACUCCAGGGA
CCACCUGGUACUGGUAAGAGUCAUUUUGCUAUUGGCCUAGCUCUCUACUACCCUUCUGCUCGCAUAGUGUAUACAGCUUGCUCUCAUGCCGCUG
UUGAUGCACUAUGUGAGAAGGCAUUAAAAUAUUUGCCUAUAGAUAAAUGUAGUAGAAUUAUACCUGCACGUGCUCGUGUAGAGUGUUUUGAUAA
AUUCAAAGUGAAUUCAACAUUAGAACAGUAUGUCUUUUGUACUGUAAAUGCAUUGCCUGAGACGACAGCAGAUAUAGUUGUCUUUGAUGAAAUU
UCAAUGGCCACAAAUUAUGAUUUGAGUGUUGUCAAUGCCAGAUUACGUGCUAAGCACUAUGUGUACAUUGGCGACCCUGCUCAAUUACCUGCAC
CACGCACAUUGCUAACUAAGGGCACACUAGAACCAGAAUAUUUCAAUUCAGUGUGUAGACUUAUGAAAACUAUAGGUCCAGACAUGUUCCUCGG
AACUUGUCGGCGUUGUCCUGCUGAAAUUGUUGACACUGUGAGUGCUUUGGUUUAUGAUAAUAAGCUUAAAGCACAUAAAGACAAAUCAGCUCAA

UGCUUUAAAAUGUUUUAUAAGGGUGUUAUCACGCAUGAUGUUUCAUCUGCAAUUAACAGGCCACAAAUAGGCGUGGUAAGAGAAUUCCUUACAC
GUAACCCUGCUUGGAGAAAAGCUGUCUUUAUUUCACCUUAUAAUUCACAGAAUGCUGUAGCCUCAAAGAUUUUGGGACUACCAACUCAAACUGU
UGAUUCAUCACAGGGCUCAGAAUAUGACUAUGUCAUAUUCACUCAAACCACUGAAACAGCUCACUCUUGUAAUGUAAACAGAUUUAAUGUUGCU
AUUACCAGAGCAAAAGUAGGCAUACUUUGCAUAAUGUCUGAUAGAGACCUUUAUGACAAGUUGCAAUUUACAAGUCUUGAAAUUCCACGUAGGA
AUGUGGCAACUUUACAA

Nsp13/helicase fragment 1 (484 nt):

GCUGUUGGGGCUUGUGUUCUUUGCAAUUCACAGACUUCAUUAAGAUGUGGUGCUUGCAUACGUAGACCAUUCUUAUGUUGUAAAUGCUGUUACG
ACCAUGUCAUAUCAACAUCACAUAAAUUAGUCUUGUCUGUUAAUCCGUAUGUUUGCAAUGCUCCAGGUUGUGAUGUCACAGAUGUGACUCAACU
UUACUUAGGAGGUAUGAGCUAUUAUUGUAAAUCACAUAAACCACCCAUUAGUUUUCCAUUGUGUGCUAAUGGACAAGUUUUUGGUUUAUAUAAA
AAUACAUGUGUUGGUAGCGAUAAUGUUACUGACUUUAAUGCAAUUGCAACAUGUGACUGGACAAAUGCUGGUGAUUACAUUUUAGCUAACACCU
GUACUGAAAGACUCAAGCUUUUUGCAGCAGAAACGCUCAAAGCUACUGAGGAGACAUUUAAACUGUCUUAUGGUAUUGCUACUGUACGUGAAGU
GCUGUCUGAC
AGAG

Nsp13/helicase fragment 3 (1296 nt):

GGUAAACCUAGACCACCACUUAACCGAAAUUAUGUCUUUACUGGUUAUCGUGUAACUAAAAACAGUAAAGUACAAAUAGGAGAGUACACCUUUG
AAAAAGGUGACUAUGGUGAUGCUGUUGUUUACCGAGGUACAACAACUUACAAAUUAAAUGUUGGUGAUUAUUUUGUGCUGACAUCACAUACAGU
AAUGCCAUUAAGUGCACCUACACUAGUGCCACAAGAGCACUAUGUUAGAAUUACUGGCUUAUACCCAACACUCAAUAUCUCAGAUGAGUUUUCU
AGCAAUGUUGCAAAUUAUCAAAAGGUUGGUAUGCAAAAGUAUUCUACACUCCAGGGACCACCUGGUACUGGUAAGAGUCAUUUUGCUAUUGGCC
UAGCUCUCUACUACCCUUCUGCUCGCAUAGUGUAUACAGCUUGCUCUCAUGCCGCUGUUGAUGCACUAUGUGAGAAGGCAUUAAAAUAUUUGCC
UAUAGAUAAAUGUAGUAGAAUUAUACCUGCACGUGCUCGUGUAGAGUGUUUUGAUAAAUUCAAAGUGAAUUCAACAUUAGAACAGUAUGUCUUU
UGUACUGUAAAUGCAUUCCUGAGACGACAGCAGAUAUAGUUGUCUUUGAUGAAAUUUCAAUGGCCACAAAUUAUGAUUUGAGUGUUGUCAAUGC
CAGAUUACGUGCUAAGCACUAUGUGUACAUUGGCGACCCUGCUCAAUUACCUGCACCACGCACAUUGCUAACUAAGGGCACACUAGAACCAGAA
UAUUUCAAUUCAGUGUGUAGACUUAUGAAAACUAUAGGUCCAGACAUGUUCCUCGGAACUUGUCGGCGUUGUCCUGCUGAAAUUGUUGACACUG
UGAGUGCUUUGGUUUAUGAUAAUAAGCUUAAAGCACAUAAAGACAAAUCAGCUCAAUGCUUUAAAAUGUUUUAUAAGGGUGUUAUCACGCAUGA
UGUUUCAUCUGCAAUUAACAGGCCACAAAUAGGCGUGGUAAGAGAAUUCCUUACACGUAACCCUGCUUGGAGAAAAGCUGUCUUUAUUUCACCU
UAUAAUUCACAGAAUGCUGUAGCCUCAAAGAUUUUGGGACUACCAACUCAAACUGUUGAUUCAUCACAGGGCUCAGAAUAUGACUAUGUCAUAU
UCACUCAAACCACUGAAACAGCUCACUCUUGUAAUGUAAACAGAUUUAAUGUUGCUAUUACCAGAGCAAAAGUAGGCAUACUUUGCAUAAUGUC
UGAUAGAGACCUUUAUGACAAGUUGCAAUUUACAAGUCUUGAAAUUCCACGUAGGAAUGUGGCAACUUUACAA

Spike S1-NTD full-length construct (937 nt):

GGACAUUUGCUUCUGACACAACUGUGUUCACUAGCAACCUCAAACAGACACCGCCACCAUGAGUCAGUGUGUUAAUCUUACAACCAGAACUCAA
UUACCCCCUGCAUACACUAAUUCUUUCACACGUGGUGUUUAUUACCCUGACAAAGUUUUCAGAUCCUCAGUUUUACAUUCAACUCAGGACUUGU
UCUUACCUUUCUUUUCCAAUGUUACUUGGUUCCAUGCUAUACAUGUCUCUGGGACCAAUGGUACUAAGAGGUUUGAUAACCCUGUCCUACCAUU
UAAUGAUGGUGUUUAUUUUGCUUCCACUGAGAAGUCUAACAUAAUAAGAGGCUGGAUUUUUGGUACUACUUUAGAUUCGAAGACCCAGUCCCUA
CUUAUUGUUAAUAACGCUACUAAUGUUGUUAUUAAAGUCUGUGAAUUUCAAUUUUGUAAUGAUCCAUUUUUGGGUGUUUAUUACCACAAAAACA
ACAAAAGUUGGAUGGAAAGUGAGUUCAGAGUUUAUUCUAGUGCGAAUAAUUGCACUUUUGAAUAUGUCUCUCAGCCUUUUCUUAUGGACCUUGA
AGGAAAACAGGGUAAUUUCAAAAAUCUUAGGGAAUUUGUGUUUAAGAAUAUUGAUGGUUAUUUUAAAAUAUAUUCUAAGCACACGCCUAUUAAU
UUAGUGCGUGAUCUCCCCUCAGGGUUUUUCGGCUUUAGAACCAUUGGUAGAUUUGCCAAUAGGUAUUAACAUCACUAGGUUUCAAACUUUACUUG
CUUUACAUAGAAGUUAUUUGACUCCUGGUGAUUCUUCUUCAGGUUGGACAGCUGGUGCUGCAGCUUAUUAUGUGGGUUAUCUUCAACCUAGGAC
UUUUCUAUUAAAAUAUAAUGAAAAUGGAACCAUUACAGAUGCUGUAGACUGUGCACUUGACCCUCUCUCAGAAACAAAGUGUACGUUGAAA

Spike S1-NTD fragment 1 (424 nt):

GGACAUUUGCUUCUGACACAACUGUGUUCACUAGCAACCUCAAACAGACACCGCCACCAUGAGUCAGUGUGUUAAUCUUACAACCAGAACUCAA
UUACCCCCUGCAUACACUAAUUCUUUCACACGUGGUGUUUAUUACCCUGACAAAGUUUUCAGAUCCUCAGUUUUACAUUCAACUCAGGACUUGU
UCUUACCUUUCUUUUCCAAUGUUACUUGGUUCCAUGCUAUACAUGUCUCUGGGACCAAUGGUACUAAGAGGUUUGAUAACCCUGUCCUACCAUU
UAAUGAUGGUGUUUAUUUUGCUUCCACUGAGAAGUCUAACAUAAUAAGAGGCUGGAUUUUUGGUACUACUUUAGAUUCGAAGACCCAGUCCCUA
CUUAUUGUUAAUAACGCUACUAAUGUUGUUAUUAAAGUCUGUGAAUUU

Spike S1-NTD fragment 3 (490 nt):

GGGUGUUUAUUACCACAAAAACAACAAAAGUUGGAUGGAAAGUGAGUUCAGAGUUUAUUCUAGUGCGAAUAAUUGCACUUUUGAAUAUGUCUCU
CAGCCUUUUCUUAUGGACCUUGAAGGAAAACAGGGUAAUUUCAAAAAUCUUAGGGAAUUUGUGUUUAAGAAUAUUGAUGGUUAUUUUAAAAUAU
AUUCUAAGCACACGCCUAUUAAUUUAGUGCGUGAUCUCCCCUCAGGGUUUUUCGGCUUUAGAACCAUUGGUAGAUUUGCCAAUAGGUAUUAACAU
CACUAGGUUUCAAACUUUACUUGCUUUACAUAGAAGUUAUUUGACUCCUGGUGAUUCUUCUUCAGGUUGGACAGCUGGUGCUGCAGCUUAUUAU
GUGGGUUAUCUUCAACCUAGGACUUUUCUAUUAAAAUAUAAUGAAAAUGGAACCAUUACAGAUGCUGUAGACUGUGCACUUGACCCUCUCUCAG
AAACAAAGUGUACGUUGAAA

A.3 Software and Computational Resources

A.3.1 Software Packages

Software	Version	Purpose
MinKNOW	24.02.26	Nanopore sequencing control
Dorado	0.5.0	Basecalling and modification detection
minimap2	2.24	Read alignment
samtools	1.15	BAM file processing
ModiDeC (epi2me-labs)	5.3.1	m^5C detection and model training
Python	3.9.x	Data analysis and visualization
NumPy	1.23.5	Numerical computing
pandas	1.5.3	Data manipulation
scikit-learn	1.2.1	Machine learning metrics
matplotlib	3.7.0	Data visualization
seaborn	0.12.2	Statistical visualization

A.3.2 Computational Resources

Component	Specification
CPU	32 cores
RAM	64 GB
GPU	NVIDIA GeForce RTX 3090 Ti (24 GB VRAM)
CUDA Version	12.2

A.3.3 Use of Artificial Intelligence and Writing Tools

Tool	Used for	Scope
Claude Code (Anthropic)	Scientific writing assistance and language editing	Throughout the entire work
Affinity Designer 2	Creation and editing of scientific figures and illustrations	Figures throughout the entire work
ChatGPT (OpenAI)	Visual modification suggestions, figure layout feedback	Selected figures
Zotero	Reference management and citation organization	Throughout the entire work

Curriculum Vitae