# ZERO-SHOT RNA MODIFICATION SITE PREDICTION WITH MOLECULAR SIDE INFORMATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

RNA modification site prediction can prioritize candidate loci for experimental validation, but site annotations are highly uneven across modification types. A modification can be chemically characterized before a sufficiently large transcriptome-wide site map is available because profiling depends on modification-specific assays and curation. Conventional sequence predictors treat modification identity as a fixed supervised label, so they cannot use this known molecular identity when the corresponding site labels are absent. We address this gap by augmenting MultiRM with standardized annotations of complete modified nucleosides and using them as molecular side information in a shared RNA-modification compatibility framework. The resulting modification embedding conditions RNA-context selection and final scoring, allowing one model to evaluate a held-out modification without a target-specific output head. Across twelve provenance-aware inductive leave-one-modification-out tasks and stricter source-separated and grouped-substructure holdouts, molecular side information improves zero-shot ranking. We further evaluate transfer to ac4C, which is not included among the twelve modification types in MultiRM. Transfer is strongest when training includes modifications with related molecular substructures and becomes less important as direct target labels accumulate. The framework therefore supports candidate-site prioritization for unseen modification types and for modifications with few annotated sites when their molecular identity is known. Resources are available at: `https://anonymous.4open.science/status/RNA_modification-1178`.

## 1 INTRODUCTION

RNA modifications alter RNA bases or ribose sugars and affect RNA processing and function (Roundtree et al., 2017; Zaccara et al., 2019). Because their effects depend on where they occur, computational site prediction can help prioritize loci for experimental validation. Yet site annotations are highly uneven across modification types: profiling requires chemistry-specific assays with different coverage and technical biases (Grozhik & Jaffrey, 2018; Grozhik et al., 2019; Acera Mateos et al., 2023). In the MultiRM source profiles, m6A has 65,178 positive sites whereas m7G has 1,036 (Song et al., 2021). A modification's molecular identity can thus be known before enough site labels are available to train a dedicated predictor.

Existing sequence predictors typically treat modification types as a fixed supervised label set. Multi-task models such as MultiRM share an RNA encoder, but identify each target through a dedicated output coordinate (Song et al., 2021). They can share sequence features across observed targets, yet cannot use a new target's molecular identity to score it without target-specific supervision. We study the resulting zero-shot problem: given an RNA window and the known chemical structure of a modification, can a model rank candidate sites when that modification contributes no site labels to training?

Molecular structure offers information that a fixed label index or parent nucleotide alone does not express. Gm and m7G share a guanosine parent, but methylation occurs at ribose 2′-O in Gm and guanine N7 in m7G. Conversely, Am and Gm share ribose 2′-O methylation across different parent bases (Figure 1a). Representing the complete modified ribonucleoside retains both distinctions. These structural relationships may provide information that transfers across modification types, but similar



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

molecular structure does not imply identical sequence preferences because which sites are modified also depends on the relevant enzymes and cellular setting.

This motivates conditioning both the RNA readout and the prediction on the queried modification. The same guanosine-centered window may provide different evidence for Gm and m7G; parent identity alone cannot specify which target is being queried. We encode molecular descriptors into a modification embedding that conditions RNA-context aggregation and a shared compatibility scorer. This allows, but does not require, different context summaries for the two queries. All learned parameters are shared across targets, so inference for a held-out modification requires its descriptor rather than a newly trained output head.

Our contributions are:

- We augment MultiRM with standardized molecular annotations of complete modified ribonucleosides and reconstruct modification-specific example provenance. The source records remain unchanged, while provenance-aware folds prevent examples selected only because of the held-out target from entering training.
- To our knowledge, we present the first sequence-based framework for zero-shot site prediction of an unseen RNA modification type from its molecular identity. A shared RNA-modification compatibility model uses the modification embedding for RNA-context aggregation and site scoring.
- We systematically evaluate transfer across twelve leave-one-modification-out tasks and stricter source-separated and grouped-substructure holdouts. We also test ac4C, which is not included in MultiRM, then vary target-specific supervision to characterize when molecular side information remains useful.

## 2 Related Work

**RNA modification site prediction and zero-shot settings.** Sequence-based RNA modification predictors range from modification-specific models such as WHISTLE to unified multi-type models such as MultiRM (Chen et al., 2019; Song et al., 2021). Rm-LR and MRM-BERT improve sequence representation with pretrained language models, but still evaluate predefined modification inventories (Liang et al., 2023; Wang & Zhou, 2024). Qiao et al. study class-incremental prediction, where labeled examples of a new modification are available (Qiao et al., 2024). AIDO.RNA provides a transferable RNA sequence encoder (Zou et al., 2024). ExpressRM studies zero-shot transfer across biological conditions for a fixed modification, whereas ORCA detects unseen modification types from nanopore direct-RNA signals (Song et al., 2025; Dong et al., 2026). These studies either retain target types seen during supervised training, introduce labels for new targets, or define zero-shot transfer along a different axis. To the best of our knowledge, this is the first sequence-based framework for zero-shot site prediction of an unseen RNA modification type from its molecular identity.

**Target-conditioned biological prediction.** Target-conditioned transfer has been studied most directly for enzyme–substrate prediction. NetPhosPan pairs phosphosite peptides with kinase information, DeepKinZero represents unseen kinases with functional and sequence attributes, and DARKIN evaluates zero-shot phosphosite–kinase association using protein-language-model representations (Fenoy et al., 2019; Deznabi et al., 2020; Sunar et al., 2025). UniZyme conditions cleavage-site prediction on enzyme and active-site information (Li et al., 2025). In these settings, side information describes an unseen protein target available at test time.

**Molecular side information.** Molecular fingerprints offer a complementary target representation when relationships are defined by chemical structure. Morgan fingerprints encode recurring molecular substructures in a fixed descriptor space (Rogers & Hahn, 2010). Modified-siRNA models such as Cm-siRPred and ENsiRNA combine sequence with chemical-modification information to predict oligonucleotide efficacy, but assume that the modification pattern is already specified (Liu et al., 2024; Tan et al., 2025). They therefore address downstream properties of a supplied design rather than site prediction for a modification type absent from site-level training labels.

# 3 METHOD

## 3.1 TASK DEFINITION

Given an RNA sequence window $x_b \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}^L$ centered at candidate site $b$ and a candidate modification $m_k$, the model predicts whether the center carries $m_k$. Binary label $y_{bk}$ is one for an annotated site and zero for its negative controls. Eligible sites are restricted to the unmodified parent nucleotide of $m_k$, removing parent-base identification as a shortcut in the primary ranking task.

For zero-shot prediction, let $\mathcal{T}$ be the set of modification types whose labels are available during training. For a held-out target $m_h \notin \mathcal{T}$, labels $y_{bh}$ are excluded from training and checkpoint selection. LOMO uses $\mathcal{T} = \{1, \ldots, K\} \setminus \{h\}$ and holds out every target in turn. The conventional fully supervised setting, in which all $K$ targets have training labels, is called the *closed-label* setting.

The distinction from a fixed-head classifier is explicit. A conventional multi-label model learns $K$ target-specific outputs $f_1(x), \ldots, f_K(x)$, and a new modification has no trained output coordinate. We instead learn a single conditional function

$$f_\theta : (x_b, q_k) \mapsto p(y_{bk} = 1), \tag{1}$$

where $q_k$ is a descriptor of modification $m_k$. Once $f_\theta$ is learned, the same function can score any modification with a valid descriptor. LOMO evaluates these scores when the held-out modification contributes no site labels to learning $\theta$.

## 3.2 MOLECULAR SIDE INFORMATION AND RNA ENCODING

Each modification is represented by a vector $q_k \in \mathbb{R}^{d_q}$ describing the complete modified ribonucleoside. A shared predictor maps $(x_b, q_k)$ to logit $s_{bk}$ and probability $p_{bk} = \sigma(s_{bk})$.

**Molecular side information.** Canonical SMILES strings are taken from MODOMICS and, when necessary, PubChem (Cappannini et al., 2024; Kim et al., 2025). The descriptor is computed from the complete modified ribonucleoside so that parent identity and modification position remain represented. RDKit produces a 2,048-bit Morgan fingerprint. Five scalar descriptors record molecular weight, topological polar surface area, hydrogen-bond donor count, hydrogen-bond acceptor count, and formal charge. A four-dimensional one-hot vector records the unmodified parent base. A shared MLP $E_{\mathrm{chem}}$ maps $q_k$ to $S_k \in \mathbb{R}^D$.

Chemical structure is supplied directly through descriptors computed before RNA-site training rather than learned from target IDs. In each LOMO fold, the descriptor encoder learns target–RNA interactions from the eleven observed modification types and applies the same mapping to the withheld descriptor. We therefore report both individual-target and grouped-holdout analyses in Section 4.

**RNA representation.** Each 51-nt window is tokenized into overlapping 3-mers and embedded using the frozen MultiRM Word2vec table (Song et al., 2021; Mikolov et al., 2013). A BiLSTM (Hochreiter & Schmidhuber, 1997) produces states $Z_b = (z_{b1}, \ldots, z_{bT})^\top \in \mathbb{R}^{T \times D}$ and center state $r_b$. The RNA encoder is shared by all modification targets. This sharing permits sequence motifs learned from abundant targets to inform a sparse or held-out target, while the chemical input determines how those shared states are interpreted.

## 3.3 TARGET-CONDITIONED COMPATIBILITY

**Target-conditioned context selection.** Multi-head cross-attention uses the modification embedding as query. For head $\ell$, $u_k^{(\ell)}$ is the projected query, and $K_b^{(\ell)}, V_b^{(\ell)} \in \mathbb{R}^{T \times d_h}$ are projected RNA keys and values:

$$\alpha_{bk}^{(\ell)} = \mathrm{softmax}\Big(K_b^{(\ell)} u_k^{(\ell)} / \sqrt{d_h}\Big), \qquad c_{bk} = W_O \,\mathrm{Concat}_{\ell=1}^{H} \left((V_b^{(\ell)})^\top \alpha_{bk}^{(\ell)}\right) + b_O. \tag{2}$$

For a fixed guanosine-centered window, queries for Gm and m7G can yield different weights and summaries $c_{b,\mathrm{Gm}}$ and $c_{b,\mathrm{m7G}}$ (Figure 1b). This formulation permits query-dependent context summaries. Independent pooling directly tests the contribution of target-conditioned RNA aggregation beyond supplying molecular information to the final scorer.

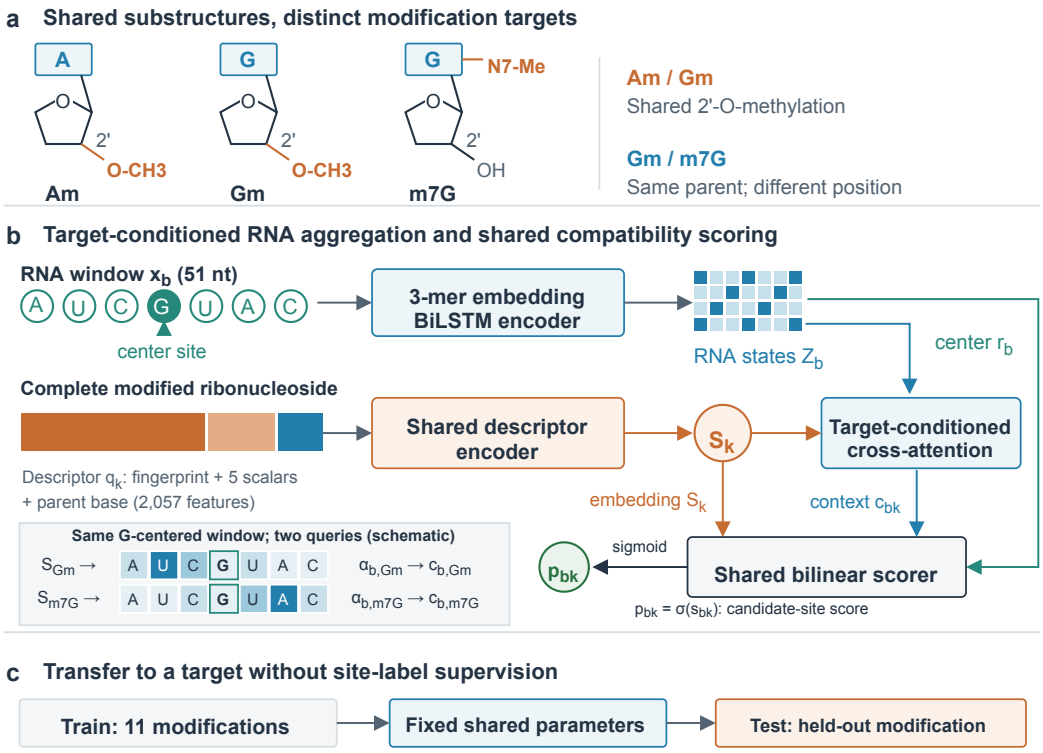



Figure 1: **Shared RNA-modification compatibility formulation.** Molecular descriptors $q_k$ of the complete modified ribonucleoside are encoded into modification embedding $S_k$, which conditions RNA-context aggregation and shared compatibility scoring. The inset illustrates possible query-dependent weighting of the same G-centered window; its colors are schematic, not measured attention. A held-out target is scored without target-specific site labels or parameters.

**Shared compatibility scoring.** The main zero-shot configuration uses bilinear compatibility,

$$s_{bk} = S_k^\top (W_r r_b + W_c c_{bk}) + \beta, \tag{3}$$

which forces prediction to depend on an interaction between the target state and RNA-derived states. This pathway asks whether the centered site and selected context are compatible with the queried modification. We use bilinear scoring for the main zero-shot analysis because it directly implements the compatibility formulation while keeping all learned parameters shared across targets. Alternative scorers test whether conclusions depend on this interaction form. Concatenation provides a flexible nonlinear matcher. A low-rank tensor scorer captures a three-way interaction between the target and the RNA states. A hypernetwork instead lets the target state modulate part of the decision function. Their parameterizations are reported in Appendix A.

### 3.4 Learning and Zero-Shot Inference

**Learning objective.** For a mini-batch of sites $\mathcal{I}$ and observed targets $\mathcal{T}$, we minimize

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{I}|\,|\mathcal{T}|} \sum_{b\in\mathcal{I}} \sum_{k\in\mathcal{T}} \left[ y_{bk} \log p_{bk} + (1 - y_{bk}) \log(1 - p_{bk}) \right]. \tag{4}$$

For each RNA window $b$, we additionally record a provenance set $P_b$ containing every modification-specific positive or negative pool from which that window entered the MultiRM union. Provenance is evaluation metadata and is never provided to the predictor. For an inductive LOMO fold with held-out target $h$, a training or validation window is eligible only when $P_b \setminus \{h\} \neq \varnothing$. Windows selected solely through the held-out target, including its target-specific negatives, are removed rather than label-masked. We also remove any remaining training or validation window whose center locus

or strand-normalized 51-nt sequence matches a held-out evaluation example. On retained windows, $y_{bh}$ is excluded from both the loss and checkpoint selection, while labels of observed modifications may still contribute.

**Zero-shot inference.** For a held-out modification $m_h$, its descriptor $q_h$ is computed by the same deterministic featurization used for observed targets and passed through the shared prediction pipeline. No target-specific parameter is added. Site labels from $m_h$ are not used to adapt the model, and inference does not substitute an observed target for $m_h$. This distinguishes the full model from nearest-target transfer and from fixed-head classifiers whose output space is defined only for supervised modifications.

**How molecular side information enters the predictor.** Molecular side information enters the predictor through two paths. First, $S_k$ changes the context summary $c_{bk}$, which makes RNA aggregation depend on the queried modification. Second, $S_k$ participates directly in Eq. 3, where it contributes to the final compatibility score. The experiments isolate these roles. Parent-only and permuted-descriptor controls change the target information while retaining target-conditioned pooling. Independent pooling keeps the full descriptor in the final scorer but removes target-dependent RNA aggregation. Its comparison with the full model therefore isolates the context-selection contribution.

# 4 Experiments

## 4.1 Benchmark Augmentation and Experimental Setup

We use the processed MultiRM benchmark and its modification-specific source collections (Song et al., 2021). MultiRM constructs positive and negative pools separately for each modification; negative sites are sampled from unmodified bases in transcripts containing positives, and each modification class is then partitioned into training, validation, and test sets. Positive sites originate from GEO and the curated RMBase and RADAR resources (Edgar et al., 2002; Xuan et al., 2018; Ramaswami & Li, 2014). Each merged example provides a 1,001-nt sequence with twelve modification labels. Following MultiRM, we use the center 51 nt. We retain the released held-out-target test assignments, but derive the primary training and validation subsets with the provenance-aware inductive filter described below.

**Molecular annotation of modification targets.** We add target-level molecular metadata without altering the source labels. For each modification, the release records the canonical structure of the complete modified ribonucleoside together with its source, the unmodified parent nucleotide, and the deterministic descriptor vector used by the model. Canonical SMILES are obtained from MODOMICS and, when necessary, PubChem. RDKit then computes the fingerprint and scalar descriptors described in Section 3.2. We also reconstruct, before merging duplicate loci or windows, the modification-specific provenance of every positive and negative pool. The release therefore specifies both molecular target annotations and the provenance needed to derive inductive LOMO and grouped holdouts. Neither provenance nor held-out labels are model inputs.

**LOMO protocol and leakage control.** For held-out target $m_h$, we first remove every training or validation example whose provenance is exclusively $m_h$. If a locus was independently collected through at least one observed modification, it may remain, but the $m_h$ label is masked. We then remove exact center-locus matches and exact strand-normalized 51-nt matches between the retained training/validation pool and the held-out $m_h$ evaluation set. The descriptor of $m_h$ is supplied only at scoring time. This is the primary provenance-aware inductive LOMO protocol. For a grouped holdout $H$, the same rule retains an example only if its provenance contains at least one target outside $H$, and labels for all targets in $H$ are masked. The study/transcript-separated experiment further groups records so that a study/transcript group cannot contribute examples to both training and evaluation. For comparison, Appendix B also reports the earlier label-masked window-retention protocol; all primary zero-shot results use provenance-aware inductive LOMO.

**Metrics and uncertainty.** The primary endpoint is targeted binary AUC ($\mathrm{AUC_b}$), which ranks positives against negative candidates with the same parent nucleotide. This removes the trivial advantage available from parent-base identity. Multi-label AUC ($\mathrm{AUC_m}$) is reported separately. The

Table 1: **Provenance-aware inductive zero-shot transfer across twelve held-out modifications.** Values are mean ± standard deviation of seed-level macro scores; inferential uncertainty across held-out modification types is reported separately. $\mathrm{AUC_b}$ is the primary metric because negatives are matched by parent nucleotide. $\mathrm{AUC_m}$ additionally includes incompatible-parent negatives.

| Model | Target information | RNA readout | Macro $\mathrm{AUC_b}$ | Macro $\mathrm{AUC_m}$ |
|---|---|---|---|---|
| Shared RNA | Common query | Target-agnostic | $0.624 \pm 0.004$ | $0.849 \pm 0.006$ |
| Parent-only | Parent identity | Target-conditioned | $0.628 \pm 0.005$ | $0.895 \pm 0.004$ |
| Permuted descriptors | Shuffled descriptors | Target-conditioned | $0.629 \pm 0.005$ | $0.880 \pm 0.005$ |
| Nearest-target transfer | Closest observed descriptor | Target-conditioned | $0.634 \pm 0.005$ | $0.905 \pm 0.004$ |
| Independent pooling | Full descriptors | Target-agnostic | $0.638 \pm 0.004$ | $0.914 \pm 0.003$ |
| Full molecular descriptors | Full descriptors | Target-conditioned | $\mathbf{0.647 \pm 0.004}$ | $\mathbf{0.924 \pm 0.003}$ |

all-target protocol uses twelve inductive LOMO folds and five matched seeds per configuration. For each target, we average the five matched-seed scores before treating the held-out modification type as the inferential unit. Main paired 95% confidence intervals are obtained by resampling the twelve target-level paired differences with replacement for 10,000 bootstrap replicates. Seed-level macro standard deviations are reported separately to describe optimization variability and are not used to quantify variation across held-out modification types. We also report the number of targets with positive seed-averaged gain and leave-one-target-out macro gains to detect effects dominated by a single modification type.

**Baseline taxonomy.** Shared RNA provides a target-agnostic reference, parent-only retains only nucleotide identity, and permuted descriptors break target–descriptor correspondence. Independent pooling removes target-dependent RNA aggregation while retaining the descriptor at the scorer, whereas nearest-target transfer reuses the most similar observed descriptor. The fixed-head model is shown only where its inability to score an unseen target is informative.

**Additional analysis protocols.** Chemical similarity is the maximum Morgan-fingerprint Tanimoto similarity between each held-out target and the observed set; its association with gain over parent-only is summarized descriptively by Spearman correlation. Low-label experiments reveal nested, class-stratified fractions of the target training partition while keeping the target excluded from checkpoint selection. The AIDO.RNA experiment uses the same target-conditioning pathway and data partitions as the scratch encoder (Appendix G).

**External ac4C evaluation.** We use two public human base-resolution ac4C datasets that are independent of MultiRM. The first is HeLa RedaC:T-seq (GEO GSE162043) (Arango et al., 2022; NCBI Gene Expression Omnibus, 2020); the second is chemical ac4C-seq in human embryonic stem cells (GEO GSE226747) (Liu et al., 2023; NCBI Gene Expression Omnibus, 2023). ac4C is absent from the twelve MultiRM targets, so no ac4C site label is used for training, model selection, or adaptation. Its complete N4-acetylcytidine structure is featurized with the same descriptor pipeline as the training targets. Positive sites are taken from the released base-resolution calls. We extract strand-aware 51-nt reference windows centered on cytidines and construct fixed parent-matched negative sets from covered cytidines in the same transcripts. The primary external metrics are $\mathrm{AUC_b}$ and AUPRC. We repeat the evaluation after exact-window removal and after a stricter 0.90 sequence-similarity filter against all MultiRM training windows. Dataset-specific construction details are reported in Appendix H.

### 4.2 Zero-Shot Prediction of Unseen Modifications

To test whether molecular side information supports prediction when an entire modification type contributes no site labels, we evaluate twelve inductive leave-one-modification-out tasks with matched data partitions and training budgets. Controls remove or corrupt molecular target information, substitute the nearest observed descriptor, or remove target-conditioned RNA aggregation, isolating which components contribute to zero-shot transfer.

Table 2: **Molecular descriptor ablation in LOMO.** All rows use the same downstream architecture.

| Representation | Macro $\mathrm{AUC_b}$ | Representation | Macro $\mathrm{AUC_b}$ |
|---|---|---|---|
| Full descriptor | $\mathbf{0.647}$ | Fingerprint + parent | $0.645$ |
| Scalars + parent | $0.635$ | Parent only | $0.628$ |
| Permuted descriptor | $0.629$ | Shared RNA | $0.624$ |

The control ordering in Table 1 separates several competing explanations for the improvement. Parent-only and permuted descriptors behave similarly, arguing against a gain from target-vector dimensionality alone. Nearest-target transfer recovers only part of the improvement, so the shared model is using information beyond copying a single observed analogue. The gap between independent and target-conditioned pooling further indicates that molecular identity contributes through RNA-context selection as well as final scoring. At the target level, gains are positive for 10 of 12 held-out modifications. The primary full-minus-parent effect is $+0.019$ $\mathrm{AUC_b}$ with a target-level paired bootstrap 95% CI of $[+0.010, +0.029]$; all twelve leave-one-target-out macro gains remain positive (Appendix B).

### 4.3 Molecular Descriptor Ablation

To identify which parts of the molecular annotation carry transferable signal, we ablate the fingerprint and the global physicochemical descriptors in matched models. Parent-only conditioning provides the reduced-information reference. The downstream architecture, data folds, and training budget remain unchanged. These comparisons distinguish substructure-level information from coarse molecular properties and parent identity.

Table 2 localizes most of the transferable signal to molecular substructure. The fingerprint-retaining variant stays close to the full model, whereas removing the fingerprint moves performance toward the reduced-information controls. Together with the descriptor-permutation control, this pattern indicates that the gain depends on correctly aligned substructure information. Coarse physicochemical summaries and parent identity alone do not account for the effect. Extra target-vector capacity is also insufficient when descriptor identity is shuffled. The target-level pattern is consistent with this interpretation. The largest gains concentrate among the four ribose 2′-O-methylations, which span different parent nucleotides while sharing a modification motif that parent-only conditioning cannot encode (Appendix B).

### 4.4 Target-Conditioned RNA Context

To isolate the role of target-dependent RNA-context selection, we compare the full model with independent pooling. Both conditions receive the same molecular descriptor through the descriptor encoder and final compatibility scorer. Only the RNA aggregation step differs. The matched comparison is given by the independent-pooling and full-model rows of Table 1, with additional attention/scorer controls in Appendix E.

The target-conditioned model is consistently stronger under this matched comparison. Because molecular information is available at the final scorer in both conditions, the remaining difference is attributable to using the modification embedding to select surrounding RNA states. Similar behavior across alternative query-conditioned attention implementations (Appendix E) further suggests that the relevant factor is access to target-dependent context rather than a particular attention parameterization.

### 4.5 Generalization under Source-Separated and Substructure Holdouts

To test dependence on annotation sources and on structurally related modifications in training, we evaluate complementary holdouts. Study/transcript-separated LOMO changes the data partition. Grouped holdouts jointly remove the four ribose 2′-O-methylations from supervision, with a stricter variant also excluding m6Am. Within each protocol, the full molecular model and parent-only control use the same data partitions.

Source separation changes which RNA examples can appear across splits, whereas grouped holdouts remove structurally related modification types from supervision. Because the protocols use different

Table 3: **Generalization under source-separated and grouped-substructure holdouts.** Full-descriptor and parent-only models are matched within each protocol. Paired gain is the macro $\mathrm{AUC_b}$ difference.

| Protocol | Full descriptors | Parent-only | Paired gain |
|---|---|---|---|
| Provenance-aware LOMO | $\mathbf{0.647 \pm 0.004}$ | $0.628 \pm 0.005$ | $+0.019$ |
| Study/transcript-separated LOMO | $\mathbf{0.608 \pm 0.006}$ | $0.598 \pm 0.006$ | $+0.010$ |
| Joint Am/Cm/Gm/Um holdout | $\mathbf{0.631 \pm 0.007}$ | $0.624 \pm 0.007$ | $+0.007$ |
| Strict substructure holdout[†] | $\mathbf{0.615 \pm 0.008}$ | $0.612 \pm 0.008$ | $+0.003$ |

[†]The four ribose-only methylations and m6Am are jointly unavailable as supervised targets.

Table 4: **External zero-shot transfer to ac4C.** ac4C is absent from MultiRM and contributes no site labels during training or model selection. AUPRC uses ten parent-matched cytidine negatives per positive.

| | HeLa RedaC:T-seq | | hESC chemical ac4C-seq | |
|---|---|---|---|---|
| Model | $\mathrm{AUC_b}$ | AUPRC | $\mathrm{AUC_b}$ | AUPRC |
| Shared RNA | $0.603 \pm 0.006$ | $0.142 \pm 0.006$ | $0.596 \pm 0.007$ | $0.136 \pm 0.006$ |
| Parent-only | $0.607 \pm 0.006$ | $0.149 \pm 0.006$ | $0.602 \pm 0.007$ | $0.143 \pm 0.006$ |
| Nearest-target transfer | $0.614 \pm 0.006$ | $0.160 \pm 0.007$ | $0.609 \pm 0.007$ | $0.153 \pm 0.006$ |
| Independent pooling | $0.619 \pm 0.005$ | $0.169 \pm 0.007$ | $0.614 \pm 0.006$ | $0.160 \pm 0.006$ |
| Full molecular descriptors | $\mathbf{0.628 \pm 0.005}$ | $\mathbf{0.181 \pm 0.008}$ | $\mathbf{0.622 \pm 0.006}$ | $\mathbf{0.171 \pm 0.007}$ |

evaluation targets and examples, each is interpreted through its matched full-versus-parent comparison. The strict substructure holdout retains a positive full-minus-parent difference after closely related modification types are removed from supervision.

### 4.6 External Zero-Shot Transfer to ac4C

The previous experiments hold out modification types included in the main benchmark. We next evaluate ac4C, which is not included in MultiRM and has independent human base-resolution site maps from HeLa RedaC:T-seq (GEO GSE162043) and hESC chemical ac4C-seq (GEO GSE226747) (Arango et al., 2022; Liu et al., 2023). Models are trained only on the twelve modification types in MultiRM. At test time, they receive the ac4C descriptor and are evaluated on both external datasets without fine-tuning or ac4C-specific checkpoint selection.

Table 4 tests the central zero-shot claim outside the twelve-target benchmark. The full model outperforms parent-only, nearest-target reuse, and target-agnostic pooling in both datasets. The same ordering across two cell types and two mapping protocols supports transfer to a modification that was absent from training and specified only by its molecular descriptor. Appendix H reports exact-window removal and a stricter 0.90 sequence-similarity filter against MultiRM training windows.

### 4.7 Chemical Similarity and Target-Level Transfer

To test whether the same relationship appears across individual held-out targets, we relate each target's gain over parent-only conditioning to its maximum fingerprint similarity to the observed modification set, using the analysis protocol defined in Section 4.1. Target-level zero-shot gains are reported in Appendix B.

Across the twelve held-out modification types, maximum fingerprint similarity is positively associated with gain over parent-only conditioning (Spearman $\rho_s = 0.52$). This pattern agrees with the grouped holdouts in Table 3: targets with closer observed molecular analogues tend to benefit more, and the gain persists after structurally related modifications are removed from supervision. Together, the two analyses characterize how the availability of related molecular substructures affects transfer.

### 4.8 Transfer across Target-Supervision Levels

To determine whether molecular side information remains useful once direct labels for the target become available, we progressively reveal nested fractions of the held-out target's training labels

Table 5: **Effect of target-specific label availability.** Values are seed-level macro-$\mathrm{AUC_b}$ gains (mean $\pm$ standard deviation) of full molecular conditioning over two matched controls.

| Target-label fraction | Gain over shared RNA | Gain over parent-only |
|---|---|---|
| 0% | $+0.023 \pm 0.004$ | $+0.019 \pm 0.004$ |
| 1% | $+0.020 \pm 0.004$ | $+0.018 \pm 0.004$ |
| 5% | $+0.016 \pm 0.004$ | $+0.014 \pm 0.004$ |
| 10% | $+0.011 \pm 0.003$ | $+0.010 \pm 0.003$ |
| 25% | $+0.006 \pm 0.003$ | $+0.005 \pm 0.003$ |
| 100% | $+0.002 \pm 0.002$ | $+0.002 \pm 0.002$ |

while keeping its test set unchanged. At every supervision level, the full model is compared with matched shared-RNA and parent-only controls. The sampling and checkpoint protocol is specified in Section 4.1.

Table 5 shows a rapid contraction of the molecular advantage as direct target supervision increases, with the same trend relative to both reference models. This pattern places the main value of molecular side information in the zero-shot and low-label regime, where it can transfer information from other modification types before enough labeled sites are available to learn target-specific sequence patterns. The same contraction is observed across target-level analyses, supporting the interpretation that molecular side information is most useful before substantial target-specific supervision is available.

### 4.9 Robustness to the RNA Encoder

To test whether transfer persists with a stronger pretrained RNA representation, we replace the BiLSTM branch with AIDO.RNA while keeping the target-conditioning pathway and data partitions fixed. The compatibility scorer and checkpoint rule are unchanged. Exact checkpoints and matched scores are reported in Appendix G.

AIDO.RNA improves absolute performance for both the shared-RNA control and the molecularly conditioned model. The full model reaches a macro $\mathrm{AUC}_b$ of 0.659, compared with 0.638 for the matched shared-RNA control, corresponding to a gain of 0.021. This gain is close to the 0.023 improvement obtained with the BiLSTM encoder. Thus, the benefit of molecular side information is retained after substantially strengthening the RNA representation. The zero-shot improvement therefore cannot be attributed only to limitations of the BiLSTM encoder or to molecular descriptors compensating for a weak sequence representation. These results also indicate that the target-conditioning mechanism can be combined with a pretrained RNA encoder without changing the basic transfer pattern observed in the main experiments.

## 5 Conclusion

We formulate RNA modification site prediction as a shared conditional problem in which molecular side information specifies the queried modification. Across provenance-aware inductive holdouts and an external unseen-modification test, matched controls indicate that transferable signal comes mainly from molecular substructure. Gains are largest when structurally related modifications are available during training and diminish as labels for the queried modification accumulate, supporting candidate-site prioritization when site annotations are scarce.

One limitation is that reliable site-level maps are currently available for only a limited set of modification types, which constrains evaluation of transfer to structurally distinct modification families. Future work will extend evaluation to families whose defining molecular substructures are absent from training as additional high-confidence site maps become available. A second limitation is that the model uses reference sequence and modification identity but no cell-type- or condition-specific information. Future work will incorporate matched biological-state inputs to support site prediction across cell types and biological conditions.

### AI USE STATEMENT

Generative AI tools assisted with language editing. The authors are responsible for the final content.

### ETHICS STATEMENT

This work uses previously released computational benchmark data and does not involve new human-subject recruitment or biological experiments. Predicted sites are hypotheses for experimental follow-up and are not clinical evidence.

### REPRODUCIBILITY STATEMENT

Section 3 specifies the molecular descriptors and both model encoders. Appendix I defines the released molecular-annotation and example-provenance schema, while Appendix J reports numerical settings. Target-level transfer results appear in Appendix B. Appendix H gives the complete external ac4C data-construction and sequence-leakage protocol. The remaining appendices define the matched descriptor, supervision, and backbone controls. In primary zero-shot evaluations, examples whose inclusion is attributable only to the held-out target are removed from training and validation, held-out labels are excluded from the loss and checkpoint selection, and exact evaluation-window overlaps are removed. Section 4.1 specifies the benchmark augmentation and evaluation protocol.

## REFERENCES


Pablo Acera Mateos, You Zhou, Kathi Zarnack, and Eduardo Eyras. Concepts and methods for transcriptome-wide prediction of chemical messenger RNA modifications with machine learning. *Briefings in Bioinformatics*, 24(3):bbad163, 2023. doi: 10.1093/bib/bbad163.

Daniel Arango, David Sturgill, Renbin Yang, Tapan Kanai, Paulina Bauer, Jyoti Roy, Ziqiu Wang, Masaki Hosogane, Sarah Schiffers, and Shalini Oberdoerffer. Direct epitranscriptomic regulation of mammalian translation initiation through n4-acetylcytidine. *Molecular Cell*, 82(15):2797–2814.e11, 2022. doi: 10.1016/j.molcel.2022.05.016.

Andrea Cappannini, Angana Ray, Elżbieta Purta, Sunandan Mukherjee, Pietro Boccaletto, S. Naeim Moafinejad, Antony Lechner, Charles Barchet, Bruno P. Klaholz, Filip Stefaniak, and Janusz M. Bujnicki. MODOMICS: a database of RNA modifications and related information. 2023 update. *Nucleic Acids Research*, 52(D1):D239–D244, 2024. doi: 10.1093/nar/gkad1083.

Kunqi Chen, Zhen Wei, Qing Zhang, Xiangyu Wu, Rong Rong, Zhiliang Lu, Jionglong Su, João Pedro de Magalhães, Daniel J. Rigden, and Jia Meng. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. *Nucleic Acids Research*, 47(7):e41, 2019. doi: 10.1093/nar/gkz074.

Iman Deznabi, Busra Arabaci, Mehmet Koyutürk, and Öznur Taştan. DeepKinZero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases. *Bioinformatics*, 36(12):3652–3661, 2020. doi: 10.1093/bioinformatics/btaa013.

Han Dong, Yongsheng Gao, Zhengyi Cai, Yi Li, Xing Li, Fangqing Zhao, and Jinyang Zhang. Comprehensive mapping of RNA modification dynamics and crosstalk via deep learning and nanopore direct RNA-sequencing. *Nature Communications*, 17:1722, 2026. doi: 10.1038/s41467-026-68419-y.

Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002. doi: 10.1093/nar/30.1.207.

Emilio Fenoy, Jose M. G. Izarzugaza, Vanessa Jurtz, Søren Brunak, and Morten Nielsen. A generic deep convolutional neural network framework for prediction of receptor–ligand interactions—NetPhosPan: application to kinase phosphorylation prediction. *Bioinformatics*, 35(7):1098–1107, 2019. doi: 10.1093/bioinformatics/bty715.

Anya V. Grozhik and Samie R. Jaffrey. Distinguishing RNA modifications from noise in epitranscriptome maps. *Nature Reviews Genetics*, 19(6):341–353, 2018. doi: 10.1038/s41576-018-0007-z.

Anya V. Grozhik, Anthony O. Olarerin-George, Miriam Sindelar, Xing Li, Steven S. Gross, and Samie R. Jaffrey. Antibody cross-reactivity accounts for widespread appearance of m1a in 5'utrs. *Nature Communications*, 10:5126, 2019. doi: 10.1038/s41467-019-13146-w.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 2025. doi: 10.1093/nar/gkae1059.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Chenao Li, Shuo Yan, and Enyan Dai. UniZyme: A unified protein cleavage site predictor enhanced with enzyme active-site knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/1ac3030fc57850b0fb11dfe9d4880ad7-Abstract-Conference.html.

Sirui Liang, Yanxi Zhao, Junru Jin, Jianbo Qiao, Ding Wang, Yu Wang, and Leyi Wei. Rm-LR: A long-range-based deep learning model for predicting multiple types of RNA modifications. *Computers in Biology and Medicine*, 164:107238, 2023. doi: 10.1016/j.compbiomed.2023.107238.

Rucong Liu, Zibaguli Wubulikasimu, Runze Cai, Fanyi Meng, Qinghua Cui, Yuan Zhou, and Yang Li. NAT10-mediated N4-acetylcytidine mRNA modification regulates self-renewal in human embryonic stem cells. *Nucleic Acids Research*, 51(16):8514–8531, 2023. doi: 10.1093/nar/gkad628.

Tianyuan Liu, Junyang Huang, Delun Luo, Liping Ren, Lin Ning, Jian Huang, Hao Lin, and Yang Zhang. Cm-siRPred: Predicting chemically modified siRNA efficiency based on multi-view learning strategy. *International Journal of Biological Macromolecules*, 264:130638, 2024. doi: 10.1016/j.ijbiomac.2024.130638.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

NCBI Gene Expression Omnibus. GSE162043: Redac:t-seq base-resolution mapping of n4-acetylcytidine in hela cells. Gene Expression Omnibus, accession GSE162043, 2020. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162043. Public functional genomics data repository.

NCBI Gene Expression Omnibus. GSE226747: Chemical ac4c-seq in human embryonic stem cells. Gene Expression Omnibus, accession GSE226747, 2023. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE226747. Public functional genomics data repository.

Jianbo Qiao, Junru Jin, Haoqing Yu, and Leyi Wei. Towards retraining-free RNA modification prediction with incremental learning. *Information Sciences*, 660:120105, 2024. doi: 10.1016/j.ins.2024.120105.

Gokul Ramaswami and Jin Billy Li. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, 42(D1):D109–D113, 2014. doi: 10.1093/nar/gkt996.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic RNA modifications in gene expression regulation. *Cell*, 169(7):1187–1200, 2017. doi: 10.1016/j.cell.2017.05.045.

Yiyou Song, Bowen Song, Daiyun Huang, Anh Nguyen, Lihong Hu, Jia Meng, and Yue Wang. Multimodal zero-shot learning of previously unseen epitranscriptomes from RNA-seq data. *Briefings in Bioinformatics*, 26(4):bbaf332, 2025. doi: 10.1093/bib/bbaf332.

Zitao Song, Daiyun Huang, Bowen Song, Kunqi Chen, Yiyou Song, Gang Liu, Jionglong Su, João Pedro de Magalhães, Daniel J. Rigden, and Jia Meng. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nature Communications*, 12:4011, 2021. doi: 10.1038/s41467-021-24313-3.

Emine Ayşe Sunar, Zeynep Işık, Mert Pekey, Ramazan Gökberk Cinbiş, and Oznur Tastan. DARKIN: a zero-shot benchmark for phosphosite–dark kinase association using protein language models. *Bioinformatics*, 41(11):btaf480, 2025. doi: 10.1093/bioinformatics/btaf480.

Wenchong Tan, Mingshu Dai, Shimin Ye, Xin Tang, Dawei Jiang, Dong Chen, and Hongli Du. ENsiRNA: A multimodality method for siRNA-mRNA and modified siRNA efficacy prediction based on geometric graph neural network. *Journal of Molecular Biology*, 437(12):169131, 2025. doi: 10.1016/j.jmb.2025.169131.

Linshu Wang and Yuan Zhou. MRM-BERT: a novel deep neural network predictor of multiple RNA modifications by fusing BERT representation and sequence features. *RNA Biology*, 21(1):1–10, 2024. doi: 10.1080/15476286.2024.2315384.

Jun-Jie Xuan, Wen-Ju Sun, Pei-Hua Lin, Kang-Rui Zhou, Sheng Liu, Li-Ling Zheng, Liang-Hu Qu, and Jian-Hua Yang. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Research*, 46(D1):D327–D334, 2018. doi: 10.1093/nar/gkx934.

Sara Zaccara, Renate J. Ries, and Samie R. Jaffrey. Reading, writing and erasing mRNA methylation. *Nature Reviews Molecular Cell Biology*, 20(10):608–624, 2019. doi: 10.1038/s41580-019-0168-5.

Shuxian Zou, Tianhua Tao, Sazan Mahbub, Caleb N. Ellington, Robin Jonathan Algayres, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P. Xing. A large-scale foundation model for RNA function and structure prediction. *bioRxiv*, 2024. doi: 10.1101/2024.11.28.625345.

## A Model Variants

The variants below test distinct forms of RNA–modification interaction. The concatenation model uses a shared nonlinear matcher. Bilinear scoring imposes direct target-RNA compatibility. The low-rank scorer adds a three-way interaction between the center, selected context, and modification embedding. The hypernetwork allows the modification embedding to change part of the decision function. These variants provide matched comparisons of the interaction form.

The additive-attention concatenation model uses

$$\alpha_{bki} = \frac{\exp\big(v_a^\top \tanh(W_z z_{bi} + W_s S_k + b_a)\big)}{\sum_{j=1}^{T} \exp(v_a^\top \tanh(W_z z_{bj} + W_s S_k + b_a))}, \qquad c_{bk} = \sum_{i=1}^{T} \alpha_{bki} z_{bi}. \tag{5}$$

Its shared MLP receives $[r_b, c_{bk}, S_k, r_b \odot S_k, c_{bk} \odot S_k]$. The low-rank scorer is

$$s_{bk} = \sum_{j=1}^{R} (r_b^\top u_j)(c_{bk}^\top v_j)(S_k^\top w_j) + \beta. \tag{6}$$

The hypernetwork projects $[r_b, c_{bk}]$ and generates target-conditioned parameters from $S_k$.

Table 6: Model variants.

| Variant | Target representation | RNA readout | Compatibility |
|---|---|---|---|
| Sequence-only | Fixed output coordinate | Sequence branch | Per-label output |
| Concat-MLP | Full descriptor | Additive attention | Shared MLP |
| Bilinear | Full descriptor | 8-head cross-attention | Eq. 3 |
| Low-rank | Full descriptor | 8-head cross-attention | Three-way factorization |
| Hypernetwork | Full descriptor | 8-head cross-attention | Generated layer |

## B Target-Level Zero-Shot Results

Table 7 decomposes the full-minus-parent comparison by held-out modification. The mean gain is $+0.019$, with positive gains for ten of twelve targets. The largest gains occur for m7G and the ribose 2′-O-methylated modifications; m5C and inosine show $-0.008$ and $-0.005$, respectively.

Table 7: **Target-level gains of full molecular descriptors over parent-only conditioning.** The mean gain is $+0.019$, and 10 of 12 held-out targets improve.

| Target | Am | Cm | Gm | Um | m1A | m5C |
|---|---|---|---|---|---|---|
| $\Delta\mathrm{AUC_b}$ | +0.038 | +0.028 | +0.034 | +0.031 | +0.014 | -0.008 |
| Target | m5U | m6A | m6Am | m7G | Ψ | I |
| $\Delta\mathrm{AUC_b}$ | +0.010 | +0.013 | +0.022 | +0.049 | +0.006 | -0.005 |

The paired target-level bootstrap gives a 95% confidence interval of $[+0.010, +0.029]$ for the full-minus-parent macro gain. Across the twelve leave-one-target-out aggregations, the gain ranges from $+0.017$ to $+0.022$. Seed-level variation is retained as an optimization-stability summary rather than the inferential unit across held-out modification types.

Table 8: **Effect of provenance-aware filtering on the zero-shot estimate.** The label-masked window-retention setting is included for comparison with the provenance-aware inductive protocol used in the primary zero-shot results.

| Protocol | Full descriptors | Parent-only | Gain |
|---|---|---|---|
| Label-masked window retention (diagnostic) | 0.661 | 0.631 | +0.030 |
| Provenance-aware inductive LOMO (primary) | **0.647** | 0.628 | +0.019 |

Table 9: **Recorded m7G window-retention diagnostic.** These values come from the earlier label-masked protocol; the primary inductive evaluation uses the provenance-aware protocol. $\mathrm{AUC_b}$ uses guanosine-matched negatives; $\mathrm{AUC_m}$ additionally includes incompatible parent nucleotides.

| Model | $\mathrm{AUC_b}$ | $\mathrm{AUC_m}$ |
|---|---|---|
| Fixed-head reference | 0.5308 | 0.4310 |
| Concat-MLP | 0.6816 | **0.9462** |
| Bilinear | **0.7176** | 0.9430 |

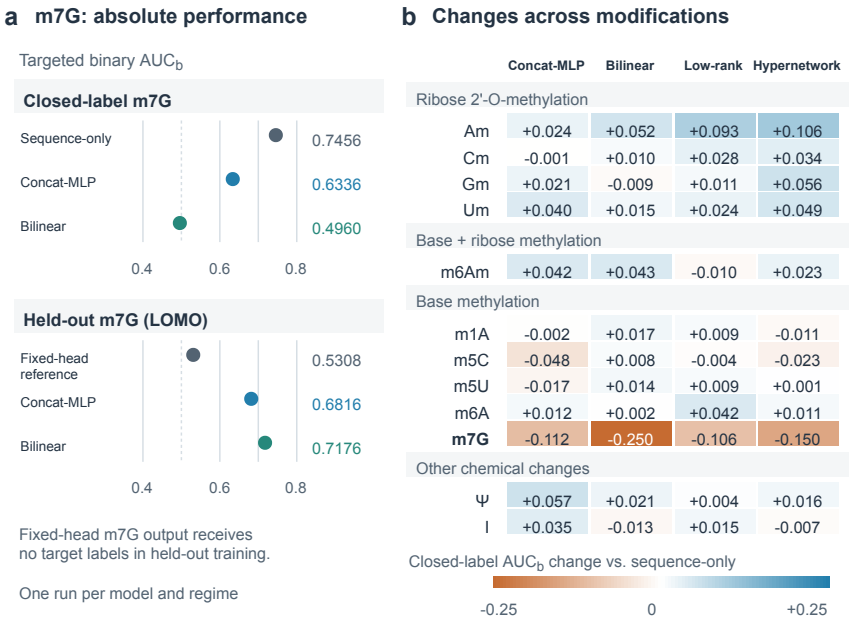



Figure 2: **Recorded supervision-regime diagnostic.** (a) m7G targeted binary AUC in the directly supervised setting and in the earlier label-masked window-retention holdout. A-MLP denotes additive-attention concatenation and M-Bilin. denotes bilinear scoring. (b) Closed-label changes in targeted binary AUC relative to the sequence-only model. The primary inductive evaluation uses the provenance-aware protocol.

## C Per-Modification Closed-Label Results

Table 10: Per-modification $\mathrm{AUC_b}$ in the closed-label setting.

| Modification | Sequence-only | Concat-MLP | Bilinear | Low-rank | Hypernetwork |
|---|---|---|---|---|---|
| Am | 0.6612 | 0.6852 | 0.7136 | 0.7540 | **0.7668** |
| Cm | 0.8416 | 0.8408 | 0.8512 | 0.8696 | **0.8752** |
| Gm | 0.9128 | 0.9340 | 0.9040 | 0.9236 | **0.9688** |
| Um | 0.8828 | 0.9228 | 0.8980 | 0.9068 | **0.9322** |
| m1A | 0.7568 | 0.7544 | **0.7736** | 0.7656 | 0.7460 |
| m5C | 0.9096 | 0.8616 | **0.9176** | 0.9052 | 0.8864 |
| m5U | 0.9448 | 0.9276 | **0.9584** | 0.9540 | 0.9460 |
| m6A | 0.8172 | 0.8292 | 0.8188 | **0.8592** | 0.8280 |
| m6Am | 0.8764 | 0.9184 | **0.9192** | 0.8660 | 0.8996 |
| m7G | **0.7456** | 0.6336 | 0.4960 | 0.6396 | 0.5952 |
| Ψ | 0.8344 | **0.8912** | 0.8552 | 0.8380 | 0.8500 |
| I | 0.5860 | **0.6208** | 0.5728 | 0.6012 | 0.5792 |
| Mean | 0.8141 | 0.8183 | 0.8065 | **0.8236** | 0.8228 |
| Median | 0.8380 | 0.8512 | 0.8532 | **0.8626** | **0.8626** |

## D Molecular Descriptor Controls

Table 11: **Molecular descriptor ablations in LOMO.** All variants use matched RNA encoder, compatibility function, training budget, and evaluation folds.

| Descriptor / control | Macro $\mathrm{AUC_b}$ |
|---|---|
| Full descriptor | $\mathbf{0.647 \pm 0.004}$ |
| Fingerprint + parent base | $0.645 \pm 0.004$ |
| Fingerprint + scalar descriptors | $0.643 \pm 0.005$ |
| Scalar descriptors + parent base | $0.635 \pm 0.005$ |
| Parent base only | $0.628 \pm 0.005$ |
| Target-agnostic shared RNA | $0.624 \pm 0.004$ |
| Permuted target-descriptor mapping | $0.629 \pm 0.005$ |

Removing the five scalar descriptors has little effect, whereas removing fingerprint information or disrupting target-descriptor correspondence reduces most of the molecular side-information advantage. This pattern indicates that transferable substructure information, rather than target identity alone, drives the zero-shot gain.

## E Attention and Compatibility Controls

Table 12: **Crossed attention and compatibility functions in LOMO.** Both attention mechanisms receive the modification embedding. Query-independent pooling is evaluated separately.

| Attention | Concat MLP | Bilinear | Low-rank | Hypernetwork |
|---|---|---|---|---|
| Additive | 0.641 | 0.645 | 0.644 | 0.642 |
| Multi-head cross-attention | **0.642** | **0.647** | **0.645** | **0.643** |

A matched bilinear model with query-independent pooling reaches macro $\mathrm{AUC_b}$ $0.638 \pm 0.004$. The similar scores across query-conditioned attention mechanisms, together with the lower query-independent result, place the emphasis on access to target-conditioned context rather than on a specific attention implementation.

## F Effect of Target-Specific Supervision

The 0% point uses the provenance-aware inductive fold, so examples selected only through the held-out target are absent from training and validation. For each target and seed, we then sample

nested, class-stratified subsets of that target's original training-provenance examples at 1%, 5%, 10%, 25%, and 100%. The selected examples are added to the training pool and their target labels are revealed. The held-out test set is unchanged across fractions, and target labels remain excluded from checkpoint selection. Smaller fractions are subsets of larger fractions within a seed so that changes along the curve are not driven by unrelated resampling. The gain from molecular side information is

Table 13: **Effect of target-specific label availability.** Gains compare the full molecular-descriptor model with shared-RNA and parent-only controls on a fixed test set.

| Target-label fraction | Gain over shared RNA | Gain over parent-only |
|---|---|---|
| 0% | $+0.023 \pm 0.004$ | $+0.019 \pm 0.004$ |
| 1% | $+0.020 \pm 0.004$ | $+0.018 \pm 0.004$ |
| 5% | $+0.016 \pm 0.004$ | $+0.014 \pm 0.004$ |
| 10% | $+0.011 \pm 0.003$ | $+0.010 \pm 0.003$ |
| 25% | $+0.006 \pm 0.003$ | $+0.005 \pm 0.003$ |
| 100% | $+0.002 \pm 0.002$ | $+0.002 \pm 0.002$ |

largest when the held-out target contributes no training labels and narrows as direct target-specific supervision increases, approaching zero in the fully supervised setting. This trend is consistent with molecular side information acting primarily as a transferable prior in low-label regimes.

## G RNA Encoder Comparison

We repeat the matched target-head comparison with a released AIDO.RNA encoder to separate the side-information effect from backbone quality. The pretrained encoder replaces the scratch BiLSTM RNA branch. We keep the molecular descriptor and compatibility scorer fixed. The data split and training procedure are also matched across the control and descriptor-conditioned models. The encoder and prediction head are fine-tuned end to end. Pretraining raises both models while preserving a positive gain from molecular side information.

Table 14: **RNA encoder comparison with matched target heads.**

| RNA encoder | Shared RNA | Full descriptors | Gain |
|---|---|---|---|
| Scratch BiLSTM | 0.624 | **0.647** | +0.023 |
| AIDO.RNA backbone | 0.638 | **0.659** | +0.021 |

The encoder comparison uses matched data and training settings. The target descriptor and scorer are unchanged.

## H External ac4C Evaluation Protocol

**Target and public data sources.** The external target is N4-acetylcytidine (ac4C), which is absent from the MultiRM target set. Its canonical complete-nucleoside structure is obtained from PubChem and featurized with the same Morgan fingerprint, scalar descriptors, and parent-base encoding used for the other targets (Kim et al., 2025). HeLa positives come from the processed RedaC:T-seq site calls released with GEO GSE162043 (Arango et al., 2022; NCBI Gene Expression Omnibus, 2020), specifically `GSE162043_RedaCTseq_ac4C_sites.txt.gz`. hESC positives come from the processed wild-type chemical ac4C-seq calls released with GEO GSE226747 (Liu et al., 2023; NCBI Gene Expression Omnibus, 2023), specifically `GSE226747_acetylated_RNA_sites.Wild_Type.xlsx`. The corresponding NAT10-KD site file is retained for a stricter sensitivity analysis but is not used to train or tune any predictor.

**Reference sequence and positive sites.** We keep each dataset in the reference assembly used by its original processing pipeline instead of lifting both datasets to a common genome build. HeLa RedaC:T-seq was aligned to hg19, whereas the hESC ac4C-seq data were processed against hg38 with Ensembl v107 gene annotation (Arango et al., 2022; Liu et al., 2023). For every released positive

site, we extract a strand-aware 51-nt reference window from the corresponding assembly. We exclude sites that cannot be assigned unambiguously to a transcript, do not have a reference cytidine at the center, or lack 25 nt of flanking reference sequence on either side. The primary hESC positive set uses the released wild-type site list. As an additional sensitivity analysis, we restrict hESC positives to wild-type sites that are absent from the released NAT10-KD site list.

**Negative construction.** External negatives are always cytidines so that the primary ranking metric remains parent matched. For HeLa, eligible negative positions are transcript-assigned reference cytidines with at least 10 reads in the $NaBH_4$-treated wild-type RedaC:T-seq data and no released ac4C call. The 10-read requirement follows the minimum coverage used by the original RedaC:T-seq analysis (Arango et al., 2022). For hESC, we reconstruct coverage from the two wild-type treated ac4C-seq replicates using the published hg38 alignment pipeline. Eligible negatives must have at least 10 reads in each treated replicate and must be absent from the released wild-type site list. For each positive, we sample up to ten negatives from the same transcript and the same broad transcript region (5′UTR, CDS, or 3′UTR). If fewer than ten region-matched candidates are available, the remaining negatives are drawn from other eligible cytidines in the same transcript. The negative sets are sampled once before model scoring and are reused for every model and training seed.

**Training, scoring, and leakage control.** For each of the five model seeds used in the main experiments, the external checkpoint is trained on all twelve MultiRM modification targets with the same objective and checkpoint rule as the full molecular model. ac4C examples are absent from training and checkpoint selection. The resulting checkpoint is frozen and receives only the ac4C molecular descriptor when scoring either external dataset. No ac4C label is used for adaptation, threshold selection, or hyperparameter tuning. The standard external set is fixed across model seeds, so the reported standard deviation reflects training variability rather than repeated test-set resampling. We apply two nested sequence-leakage controls. The first removes exact strand-normalized 51-nt matches to any MultiRM training window. The second uses CD-HIT-EST-2D to remove every external 51-nt window with sequence identity of at least 0.90 to any MultiRM training window, following the inter-dataset de-redundancy principle used in MRM-BERT (Wang & Zhou, 2024).

**Metrics.** We report $\mathrm{AUC_b}$ on the parent-matched cytidine ranking task and AUPRC on the fixed 1:10 positive-to-negative evaluation set. AUPRC is used for matched model comparisons on the same constructed 1:10 evaluation set. We report the same metrics after exact-match removal and after the 0.90 sequence-similarity filter. We also report a separate hESC WT-specific-call sensitivity analysis in which positives are restricted to wild-type calls absent from the released NAT10-KD call list.

Table 15: **External ac4C sequence-leakage controls.** Exact-match filtering removes identical 51-nt windows; the stricter filter removes external windows with $\geq 0.90$ identity to any MultiRM training window using CD-HIT-EST-2D.

| Dataset | Filter | Parent-only $\mathrm{AUC_b}$ | Full $\mathrm{AUC_b}$ | $\Delta\mathrm{AUC_b}$ | $\Delta$AUPRC |
|---|---|---|---|---|---|
| HeLa | Exact 51-nt match | 0.604 | **0.624** | +0.020 | +0.030 |
| HeLa | $< 0.90$ similarity | 0.596 | **0.612** | +0.016 | +0.024 |
| hESC | Exact 51-nt match | 0.600 | **0.618** | +0.018 | +0.025 |
| hESC | $< 0.90$ similarity | 0.593 | **0.608** | +0.015 | +0.020 |

Table 16: **hESC WT-specific-call sensitivity analysis.** Positives are wild-type calls absent from the released NAT10-KD call list. The second row additionally applies the 0.90 similarity filter against MultiRM training windows.

| Filter | Parent-only $\mathrm{AUC_b}$ | Full $\mathrm{AUC_b}$ | $\Delta\mathrm{AUC_b}$ | $\Delta$AUPRC |
|---|---|---|---|---|
| No sequence filter | 0.598 | **0.617** | +0.019 | +0.025 |
| $< 0.90$ similarity | 0.590 | **0.606** | +0.016 | +0.021 |

# I Data Sources and Molecular Annotation Schema

The released benchmark augmentation stores molecular target metadata together with example-level provenance used only to construct evaluation folds. The underlying MultiRM labels are not altered. Table 17 summarizes the fields needed to reconstruct the molecular side information and inductive protocols.

Table 17: Released target-level molecular annotation schema.

| Field | Description |
|---|---|
| Modification ID | MultiRM target name used to join molecular metadata to site labels. |
| Complete modified ribonucleoside | Canonical molecular structure used for featurization. |
| Structure source | MODOMICS entry or PubChem fallback used to obtain the canonical structure. |
| Parent nucleotide | Unmodified RNA nucleotide associated with eligible candidate sites. |
| Molecular descriptors | 2,048-bit Morgan fingerprint, five physicochemical scalars, and parent-base one-hot vector. |
| Example provenance | Modification-specific positive/negative pool(s) through which each RNA window entered the merged benchmark; used only to construct inductive evaluation folds. |
| Evaluation metadata | LOMO target identity and membership of the source-separated and grouped-substructure holdouts. |

Table 18: MultiRM positive-site provenance.

| Modification | Positive sites | Assay/repository | Public accession(s) |
|---|---|---|---|
| m6A | 65,178 | m6A-CLIP-seq, miCLIP | GSE71154, GSE86336, GSE98623, GSE63753 |
| Ψ | 3,137 | Ψ-seq, Pseudo-seq, CeU-Seq, RBS-Seq | GSE60047, GSE58200, GSE63655, GSE90963 |
| m1A | 16,380 | miCLIP, m1A-MAP, RBS-Seq, m1A-seq | GSE97908, GSE102040, GSE90963, GSE97419, GSE70485 |
| m6Am | 2,447 | miCLIP | GSE122948, GSE78040, GSE63753 |
| Am/Cm/Gm/Um | 1,591/1,878/1,471/2,253 | Nm-seq and RMBase 2′-O-Me | GSE90164, RMBase records |
| m5C | 12,936 | bisulfite sequencing | GSE122260 |
| m7G | 1,036 | m7G-seq | GSE112276 |
| m5U | 1,696 | miCLIP and FICC-seq | GSE109183 |
| I | 52,618 | RADAR A-to-I records | RADAR records |

# J Implementation Settings

Table 19: Implementation settings for the evaluated configurations.

| Setting | Value |
|---|---|
| RNA window length $L$ | 51 nucleotides |
| Tokenization | Overlapping 3-mers ($T = 49$) |
| Token embedding dimension $d_e$ | 300 (frozen) |
| RNA and chemical dimension $D$ | 512 (256 BiLSTM units per direction) |
| Morgan fingerprint length $d_f$ | 2,048 bits |
| Chemical descriptor dimension $d_q$ | 2,057 ($d_f + 5 + 4$) |
| Attention heads $H$ | 8 (head width $d_h = 64$) |
| Low-rank factorization rank $R$ | 32 |
| Hypernetwork projection dimension $d$ | 32 |
| Optimizer | Adam (Kingma & Ba, 2015) |
| Training epochs / batch size | 50 / 128 |
| Initial learning rate | $10^{-4}$ |
| Learning-rate schedule | factor 0.8, patience 5 epochs on validation loss |