Decoding Histone Modification Signatures of Non-Coding RNAs via Foundation Models

Nishant Sharma
Department of Electrical and Computer Engineering
University of California San Diego
nishant.sharma.iitd@gmail.com

Mohammad Atif Quamar
Independent Researcher
atif7102@gmail.com

Pengtao Xie
Department of Electrical and Computer Engineering
University of California San Diego
p1xie@ucsd.edu

Abstract

Histone modifications help regulate ncRNA genes, but measuring these interactions at scale is difficult. High-throughput experimental techniques such as ChIP-seq are costly and time-consuming, limiting their scalability for mapping histone modifications across diverse cell types and histone markers. We test whether sequence alone can predict histone–ncRNA regulation by training a single marker-conditioned classifier that outputs 50 histone marks. We evaluate two inputs: (i) spliced transcript RNA sequence and (ii) genomic context comprising the gene body plus up to 30 kb upstream DNA. On a curated benchmark with Ensembl coordinates, the context-based model attains a micro-AUROC of 0.95. Despite using frozen pretrained encoders and no cell-type-specific tracks, the approach is simple and data-efficient, providing a practical baseline for studying ncRNA–histone modification interactions.

1 Introduction

Over the past few decades, research has revealed that epigenetic changes play an important role both in the functioning of normal organisms and in disease progression. They are described as mechanisms that can lead to inherited changes in phenotype or gene expression but do not involve alterations in the DNA sequence Feinberg (2007). Among the key epigenetic regulators are DNA methylation/ demethylation, chromatin remodeling, histone modifications, and ncRNAs Wei et al. (2017). Moreover, it is increasingly clear that these mechanisms do not act in isolation but rather coordinate gene expression simultaneously, combining in a wide regulatory network Bure and Nemtsova (2021); Fuso et al. (2020).

Histone modifications are covalent post-translational changes, often occurring on the N-terminal tails of histone proteins. These modifications can affect how tightly DNA is wrapped around histones by altering histone–DNA interactions or by recruiting proteins that activate or repress gene expression Millán-Zambrano et al. (2022). Common types of histone modifications include acetylation, methylation, phosphorylation, sumoylation, and ubiquitylation. They are responsible to regulate multifarious biological processes including chromosome wrapping Bannister and Kouzarides (2011); Brehove et al. (2015) transcriptional activation and de-activation Kouzarides (2007); Binder et al. (2013); Cheung et al. (2000), damaging and repairing of DNA Narlikar et al. (2002); Kristeleit et al. (2004). For instance, histone amino (N)-terminal tails modifications influence internucle-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

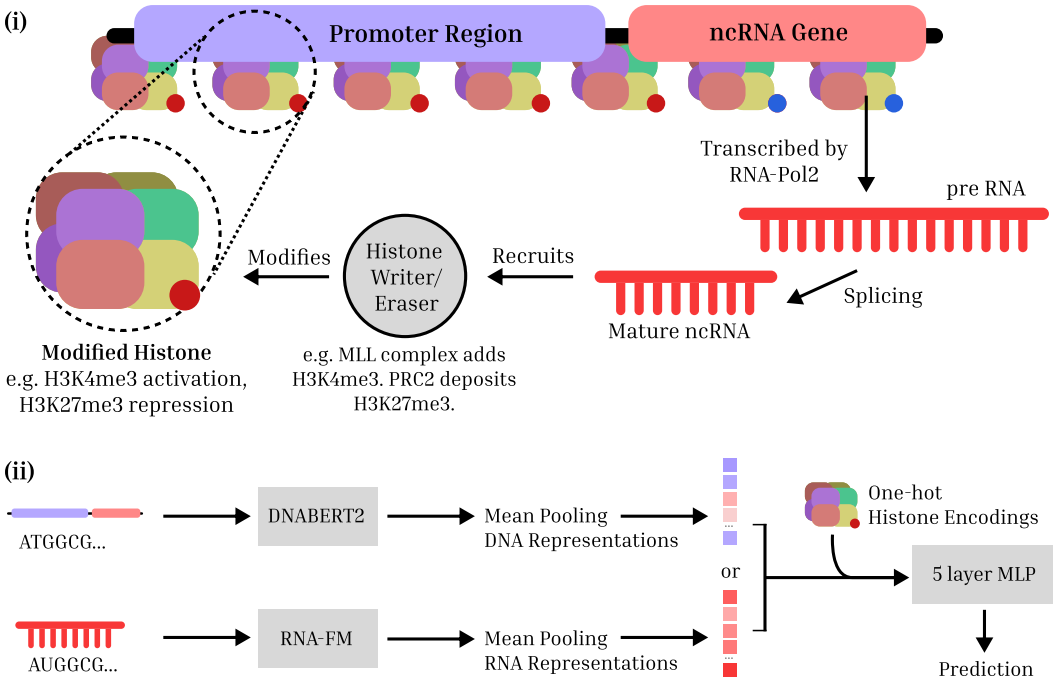


Figure 1 | (i) Illustration of a mechanism by which ncRNAs influence gene expression through histone modifications and vice versa. A transcribed ncRNA recruits histone writers or erasers (such as MLL or PRC2) to gene promoters, modulating histone marks like H3K4me3 or H3K27me3. This represents one of several known ncRNA-mediated regulatory pathways. (ii) A schematic of our method of predicting histone-modifications on ncRNA coding genes.

osomal exchanges and are capable to modify chromatin structures which ultimately affect gene expression Peterson and Laniel (2004), contributing to the development of complex diseases such as cancer. Bannister and Kouzarides (2011).

Importantly, histone modifications also regulate the transcription of non-coding RNAs (ncRNAs), adding another layer to epigenetic control (as shown in Fig. 1 (i)). In this context, understanding how specific histone marks correlate with ncRNA loci is essential to uncovering the mechanisms by which chromatin state regulates gene expression Bure et al. (2022); Sati et al. (2012). While certain histone marks are associated with active transcription, emerging evidence suggests that transcription itself may also be necessary to establish or maintain these modifications Wang et al. (2022). This reciprocal relationship reflects the complexity of epigenetic regulation involving non-coding RNAs.

Histone modifications are primarily profiled using high-throughput techniques such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) O'Geen et al. (2011); Park (2009). However, as the volume of sequencing data grows, there is an increasing need for computational frameworks capable of identifying histone modification patterns associated with ncRNA transcription.

To address this, we introduce a simple *marker-conditioned* framework for predicting 50 histone–ncRNA interactions from sequence. We evaluate two configurations: one using RNA-FM embeddings of ncRNA transcripts, and another using DNABERT2 embeddings of promoter + proximal DNA. To the best of our knowledge, this is the first systematic treatment of 50 distinct histone–ncRNA interaction tasks in a single classifier, achieving AUROC up to 0.95 on curated human/mouse data and outperforming baselines. The model is sequence-only, requiring no chromatin-accessibility or cell-type-specific tracks, and is data-efficient by leveraging frozen foundation model (FM) embeddings. This further demnstrates that adding upstream promoter context (gene + 30 kbp) *improves* accuracy and related metrics over RNA-only inputs, underscoring the value of genomic context for ncRNA regulation. Finally, we provide per-marker evaluation and analyses of learned histone-mark embeddings that recover known biochemical relationships, establishing a simple, interpretable baseline for ncRNA–chromatin prediction.

2 Method

2.1 Dataset, Model Inputs and Architecture

We start from RNAInter v4.0 and retrieve corresponding transcripts and gene coordinates from Ensembl Harrison et al. (2024) and NCBI NA et al. (2024), labeling 50 histone marks per RNA and defining a regulatory domain of 30 kb upstream of the TSS; unobserved pairs are treated as negatives. Data are split by unique RNA (10 % validation, 10 % test) with no overlap.

Table 1: Test set performance for two configurations of the same model architecture using different sequence types and pretrained embeddings. All metrics are micro-averaged across different histone modifications. Per-marker accuracies are shown in Fig. 2.

Configuration	Input Type	Embedding	Accuracy	Precision	Recall	F1-score	MCC	AUROC
RNA-FM	Spliced RNA sequence	RNA-FM	0.7425	0.6496	0.9051	0.7563	0.5298	0.86
DNABERT2	Gene + 30 kbp upstream	DNABERT2	0.8594	0.7837	0.9379	0.8539	0.7314	0.95

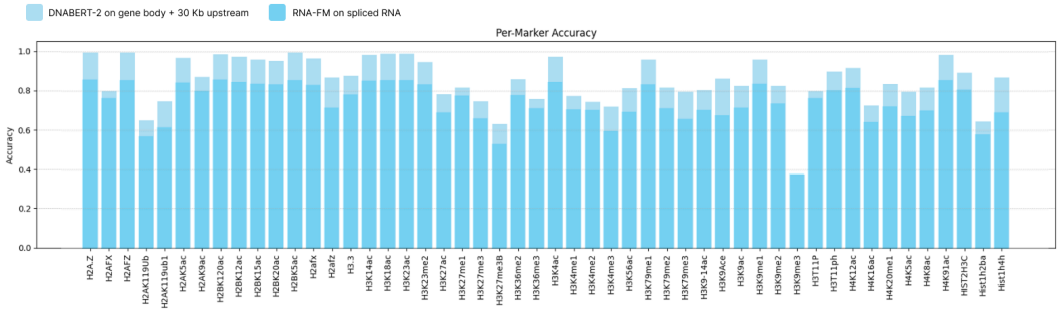


Figure 2 | Per-marker test accuracy on the two configurations. Each bar represents held-out accuracy for one of the 50 histone modifications.

Transcripts are embedded via RNA-FM (mean-pooled over fixed-length chunks), and gene (or promoter + gene) DNA sequences via DNABERT-2, with embeddings concatenated with a 50-dimensional one-hot histone-mark vector for marker conditioning. The model (see Fig. 1 (ii)) accepts two inputs: a sequence embedding and a histone marker vector. For RNA-based experiments, embeddings were generated using RNA-FM Shen et al. (2024) on spliced transcript sequences. For DNA-based settings, DNABERT2 Zhou et al. (2023) was applied either to the gene body alone or to the gene body plus a 30 kilobase pairs (kbp) upstream promoter region. Each of the 50 histone modifications was represented by a one-hot encoded vector. The embedding and marker vectors were concatenated and fed into a multilayer perceptron (MLP) with five hidden layers, totaling approximately 5M parameters Belkin et al. (2019). Detailed preprocessing, model architecture, hyperparameters, and compute details are provided in Appendices A, B, and C.

2.2 Evaluation Setup

We evaluated our model on a set of 27196 unique RNA sequences from *Homo sapiens* and *Mus musculus*, derived from the RNAInter v4.0 dataset Kang et al. (2021). Each RNA is annotated with 50 distinct histone markers using genomic data sourced from Ensembl Harrison et al. (2024). The dataset is imbalanced in terms of positive and negative class distributions across histone markers (see Extended Data Fig. 1). To handle this, we adopted class-weighted loss using positive class weights. To evaluate generalization, we held out 10% of the unique RNA sequences for testing. These sequences were never seen during training or validation. The model is both taxonomically and epigenome agnostic, relying only on raw genomic or transcript sequences. Although the dataset is mostly dominated by ncRNA, see Extended Fig. 2 for RNA category distribution.

3 Experimental Results

3.1 Accuracy and AUROC Across Histone Marks

We evaluated model performance across two input conditions: (1) spliced RNA sequences embedded with RNA-FM, and (2) gene body plus 30 kbp upstream promoter regions, embedded with DNABERT2. As shown in Table 1, the complete sequence DNABERT2 configuration achieved an overall micro-AUROC of **0.95**, followed by an overall test micro-AUROC of **0.86** on spliced RNA. We report averaged scores for precision, recall, F1-score, and Matthews Correlation Coefficient (MCC), with detailed per-marker AUROC (including baselines) shown in Extended Fig. 4, and per-marker accuracy shown in Fig. 2. The inclusion of upstream promoter regions consistently improved all metrics, supporting the hypothesis that distal regulatory elements contribute substantially to histone modification patterns Shlyueva et al. (2014); Heintzman et al. (2007).

3.2 Comparison with Baselines

We evaluated our method against DeepHistone Yin et al. (2019). Since Histone-Net Asim et al. (2023) was not open-sourced, it could not be included in our comparisons. We refer to the unmodified implementation of DeepHistone as Vanilla DeepHistone. We successfully reproduced Vanilla DeepHistone (totaling 34.4 million trainable parameters, including 5.1M in the convolutional sequence tower and 29.4M in the fully connected classification head) on RNAInter v4.0. However, the model exhibited significantly limited generalization performance in this setting. Replacing the flatten-then-dense head with a global pooling head (see Appendix D) yielded a far more compact model, parameter savings on the order of one log, and, crucially, allowed ingestion of a broader regulatory context (11 kbp promoter + 1 kbp gene body). Restricting the model's receptive field to 11 kbp truncated gene window provided the best performance with the baseline. This adapted DeepHistone achieved a macro-averaged AUROC of 0.89 and micro-averaged AUROC of 0.94 on the held-out RNAInter test fold, outperforming the spliced-ncRNA framework evaluated under identical splits.

Table 2: Test-set macro- and micro-averaged AUROC on RNAInter v4.0 for baseline, RNA-FM and DNA-BERT2. See Appendix F for evaluation metrics and D for architecture of adapted DeepHistone. All values are reported on a held-out test set.

Configuration	**Macro-AUROC**	**Micro-AUROC**
DeepHistone (Vanilla)	0.49	0.78
DeepHistone (Adapted)	0.89	0.94
RNA-FM	0.85	0.86
DNABERT2	**0.92**	**0.95**

4 Related Work

Several deep learning frameworks have been proposed to model histone modifications and their impact on gene regulation. DeepHistone Yin et al. (2019) introduced one of the earliest convolutional models for predicting histone modification states from DNA sequences and chromatin accessibility data. While effective, it relied heavily on cell-type-specific accessibility signals, which may not always be available in practice. More recently, Histone-Net Asim et al. (2023) extended this line of work by combining supervised sequence embeddings with multi-paradigm learning strategies to predict histone occupancy and modification. In contrast, our model leverages transfer learning from frozen FMs and requires significantly fewer labeled training instances while generalizing across 50 histone markers. Beyond histone-specific models, broader frameworks for epigenomic prediction have emerged. DeepChrome Singh et al. (2016) employed convolutional neural networks to predict gene expression from local histone modification signals.

5 Conclusion

We presented a simple *marker-conditioned* framework that uses RNA-FM embeddings of ncRNA transcripts and DNABERT2 embeddings from promoter + proximal DNA to predict **50** histone–ncRNA interactions using a single model. Across curated human and mouse datasets, the model achieved strong performance (micro-AUROC up to 0.95) and outperformed non-pretrained baselines, while ablations showed that adding upstream promoter context (gene + 30 kbp) consistently improves accuracy over RNA-only inputs, underscoring the regulatory value of genomic context. Leveraging frozen FM features yields a scalable, data-efficient solution that does not depend on auxiliary epigenomic assays. A key limitation is the treatment of unobserved pairs as negatives, placing the task in a positive-unlabeled regime; future work will incorporate PU-aware estimators and calibration.

References

Andrew P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447: 433–440, 2007.

Jian-Wei Wei, Kai Huang, Chao Yang, and Chun-Sheng Kang. Non-coding rnas as regulators in epigenetics. *Oncology Reports*, 37:3–9, 2017. doi: 10.3892/or.2016.5236.

Irina V Bure and Marina V Nemtsova. Methylation and noncoding rnas in gastric cancer: Everything is connected. *International Journal of Molecular Sciences*, 22(11):5683, 2021.

Andrea Fuso, Tiziana Raia, Michela Orticello, and Marco Lucarelli. The complex interplay between dna methylation and mirnas in gene expression regulation. *Biochimie*, 173:12–16, 2020. doi: 10.1016/j.biochi.2020.02.006.

Gonzalo Millán-Zambrano, Adam Burton, Andrew J. Bannister, and Robert Schneider. Histone post-translational modifications — cause and consequence of genome function. *Nature*, 23:563–580, 2022.

Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21:381–395, 2011.

Matthew Brehove, Tao Wang, Justin North, Yi Luo, Sarah J Dreher, John C Shimko, Jennifer J Ottesen, Karolin Luger, and Michael G Poirier. Histone core phosphorylation regulates dna accessibility. *J Biol Chem*, 290(37):22612–22621, 2015.

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

Hans Binder, Lydia Steiner, Jens Przybilla, Thimo Rohlf, Sonja Prohaska, and Jörg Galle. Transcriptional regulation by histone modifications: towards a theory of chromatin re-organization during stem cell differentiation. *Phys. Biol.*, 10(2):026006, 2013.

Peter Cheung, C.David Allis, and Paolo Sassone-Corsi. Signaling to chromatin through histone modifications. *Cell*, 103:263–271, 2000.

Geeta J. Narlikar, Hua Y. Fan, and Robert E. Kingston. Cooperation between complexes that regulate chromatin structure and transcription. *Cell*, 108(4):475–487, 2002.

Rebecca Kristeleit, Lucy Stimson, Paul Workman, and William Aherne. Histone modification enzymes: novel targets for cancer drugs. *Expert Opin. Emerg. Drugs*, 9(1):135–154, 2004.

Craig L. Peterson and Marc-André Laniel. Histones and histone modifications. *Curr. Biol.*, 14(14): R546–R551, 2004.

Irina V. Bure, Maria V. Nemtsova, and Ekaterina B. Kuznetsova. Histone modifications and non-coding rnas: Mutual epigenetic regulation and role in pathogenesis. *International Journal of Molecular Sciences*, 23(10):5801, 2022. doi: 10.3390/ijms23105801.

Satish Sati, Sourav Ghosh, Vaibhav Jain, Vinod Scaria, and Shantanu Sengupta. Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding rna loci. *Nucleic Acids Research*, 40(20):10018–10031, 2012. doi: 10.1093/nar/gks776.

Zhong Wang, Alexandra G. Chivu, Lauren A. Choate, Edward J. Rice, Donald C. Miller, Tinyi Chu, Shao-Pei Chou, Nicole B. Kingsley, Jessica L. Petersen, Carrie J. Finno, Rebecca R. Bellone, Douglas F. Antczak, John T. Lis, and Charles G. Danko. Prediction of histone post-translational modification patterns based on nascent transcription data. *Nature*, 54:295–305, 2022.

Henriette O'Geen, Lorigail Echipare, and Peggy J Farnham. Using chip-seq technology to generate high-resolution profiles of histone modifications. *Methods Mol Biol.*, 791:265–286, 2011.

Peter J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nature*, 10:669–680, 2009.

Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, et al. Ensembl 2024. *Nucleic Acids Research*, 52:D891–D899, 2024.

O'Leary NA, Holmes JB Cox E, Anderson WR, Falk R, Hem V, Tsuchiya MTN, Schuler GD, Zhang X, Torcivia J, Ketter A, Breen L, Cothran J, Bajwa H, Tinne J, Meric PA, Hlavina W, and Schneider VA. Exploring and retrieving sequence and metadata for species across the tree of life with ncbi datasets. *Nature, Scientific Data*, 11(732), 2024.

Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, pages 1–12, 2024.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome, 2023.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

Juanjuan Kang, Qiang Tang, Jun He, Le Li, Nianling Yang, Shuiyan Yu, Mengyao Wang, Yuchen Zhang, Jiahao Lin, Tianyu Cui, Yongfei Hu, Puwen Tan, Jun Cheng, Hailong Zheng, Dong Wang, Xi Su, Wei Chen, and Yan Huang. Rnainter v4.0: Rna interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Research*, 50(D1):D326–D332, 10 2021. doi: 10.1093/nar/gkab997.

Daria Shlyueva, Gabriel Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15:272–286, 2014. doi: 10.1038/nrg3682.

Nathan D. Heintzman, R. K. Stuart, Gary Hon, Yuchun Fu, Charles W. Ching, R. David Hawkins, Leah O. Barrera, Sanne Van Calcar, Chunxia Qu, Kenneth A. Ching, Wei Wang, Zhiping Weng, Robert D. Green, Gregory E. Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318, 2007. doi: 10.1038/ng1966.

Qijin Yin, Mengmeng Wu, Qiao Liu, Hairong Lv, and Rui Jiang. Deephistone: a deep learning approach to predicting histone modifications. *BMC Genomics*, 20(193), 2019. doi: 10.1186/s12864-019-5489-4.

Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Imran Malik, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. Histone-net: a multi-paradigm computational framework for histone occupancy and modification prediction. *Complex and Intelligent Systems*, 9:399–413, 2023.

Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 2016. doi: 10.1093/bioinformatics/btw427.

Ke-Ren Zhou, Shun Liu, Wen-Ju Sun, Ling-Ling Zheng, Hui Zhou, Jian-Hua Yang, and Liang-Hu Qu. Chipbase v2.0: Decoding transcriptional regulatory networks of non-coding rnas and protein-coding genes from chip-seq data. *Nucleic Acids Research*, 45:D43–D50, 2017.

Yunqing Lin, Tianyuan Liu, Tianyu Cui, Zhao Wang, Yuncong Zhang, Puwen Tan, Yan Huang, Jia Yu, and Dong Wang. Rnainter in 2020: Rna interactome repository with increased coverage and annotation. *Nucleic Acids Research*, 48:D189–D197, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814. Omnipress, 2010.

Lutz Prechelt. Early stopping — but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

A Dataset preprocessing

Our dataset is derived from *RNAInter v4.0*, a curated repository of RNA interactomes that includes experimentally validated interaction pairs between RNA sequences and histone modifications. From this resource, we extracted a total of 1,060,684 positive interaction pairs, where `Interactor1.Symbol` represents the RNA and `Interactor2.Symbol` denotes the associated histone modification. RNA sequences were retrieved using programmatic access to the Ensembl Harrison et al. (2024) database. Specifically, the Ensembl subset provided 27196 unique RNA identifiers linked to 734,644 interactions.

The RNA species in the dataset span a diverse range of biotypes, including long non-coding RNAs (lncRNAs), microRNAs (miRNAs), messenger RNAs (mRNAs), small nuclear RNAs (snRNAs), transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs), scaRNAs, sRNAs, ribozymes, pseudogenes, and sequences labeled as "unknown" (see Extended Data Fig. 2).

Following the dataset literature which suggests that the histone modification signals were taken from the promoter region of the gene body Zhou et al. (2017); Lin et al. (2020), we defined the regulatory domain of each RNA as the 30 kbp region upstream from its transcription start site (TSS). RNAInter provides ChIP-Seq-supported interactions, which vary in confidence but are treated uniformly in this study. To enable supervised binary classification, we constructed a complementary set of negative samples by pairing RNAs with histone modifications not observed as interactors in the dataset. This approach assumes that any RNA–histone pair not explicitly annotated constitutes a negative interaction.

To promote model generalization, we partitioned the dataset such that RNA sequences in the training, validation, and test sets are mutually exclusive. First, 10% of all RNAs were set aside as the test set. From the remaining RNAs, an additional 10% were held out as validation data. This ensures that both evaluation phases are conducted on RNA molecules never seen during training, and avoids any potential data leakage through shared sequence information.

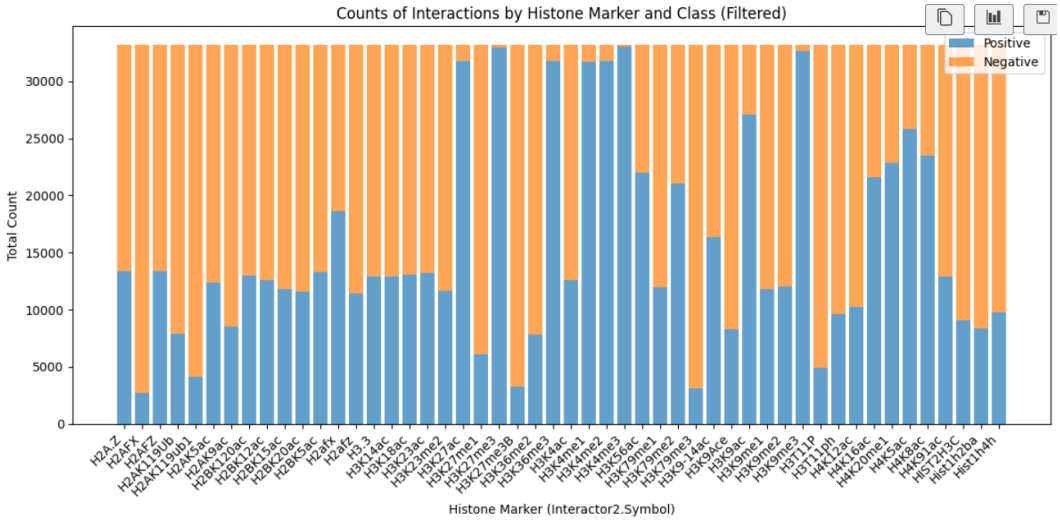
To address class imbalance, we explored two independent strategies during training. The first involved per-histone-marker balancing, where we ensured a 50:50 ratio of positive and negative samples for each marker by undersampling the overrepresented class. Although this reduced the total number of samples for some markers, it standardized the training distribution and simplified the binary classification task. In contrast, the second approach retained the full dataset and applied a pos_weight parameter in the loss function to assign greater importance to the minority class. This weighting strategy preserved data availability while guiding the model to better learn from underrepresented examples. All evaluation is done under the latter scheme.

Lastly, the dataset is epigenome-agnostic. While RNAInter includes data from both *Homo sapiens* and *Mus musculus*, our model does not stratify interactions by species, tissue, cell type, or experimental condition. All RNA–histone interactions are pooled, thereby removing epigenomic context from consideration. While this design decision facilitates broader generalization, it may obscure biological variability that depends on cellular or organism-specific chromatin states.

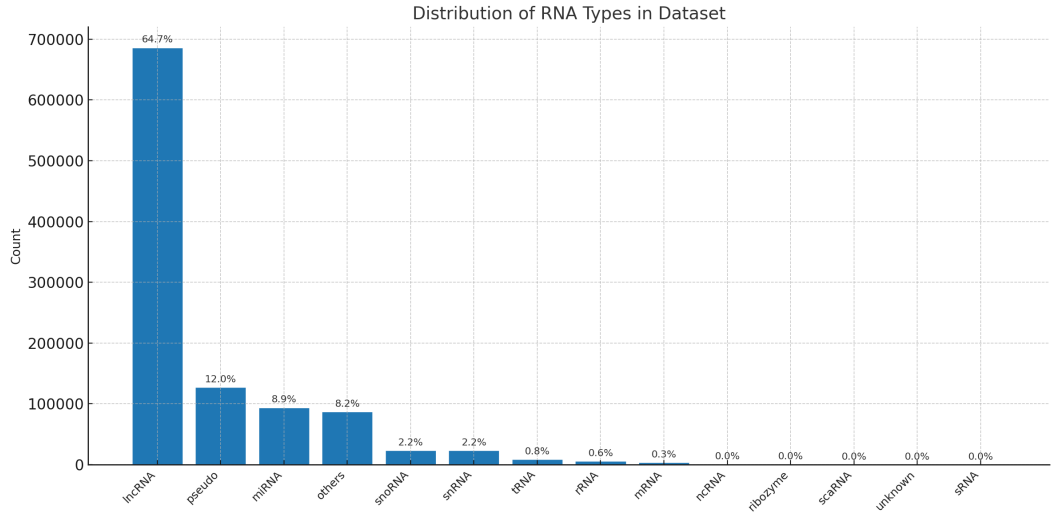
B Encoding RNA and histone modifications

To encode RNA sequences, we split each sequence into 1,000 fixed-length segments. Each segment is embedded using the RNA-FM model, a foundation model pretrained on large-scale RNA data. The segment-wise embeddings are then aggregated by computing their mean, yielding a single representation vector for the entire RNA sequence.

A similar approach is used for DNA-based models. For each RNA, we retrieve the corresponding gene or gene + promoter region, and encode the DNA sequence using DNABERT-2, a pretrained masked language model for genomic data. The sequence is likewise split into segments, and their embeddings are averaged to form a single fixed-length representation. Despite being trained on different molecular modalities, RNA-FM and DNABERT-2 produced comparably strong performance in downstream histone–RNA interaction prediction tasks.

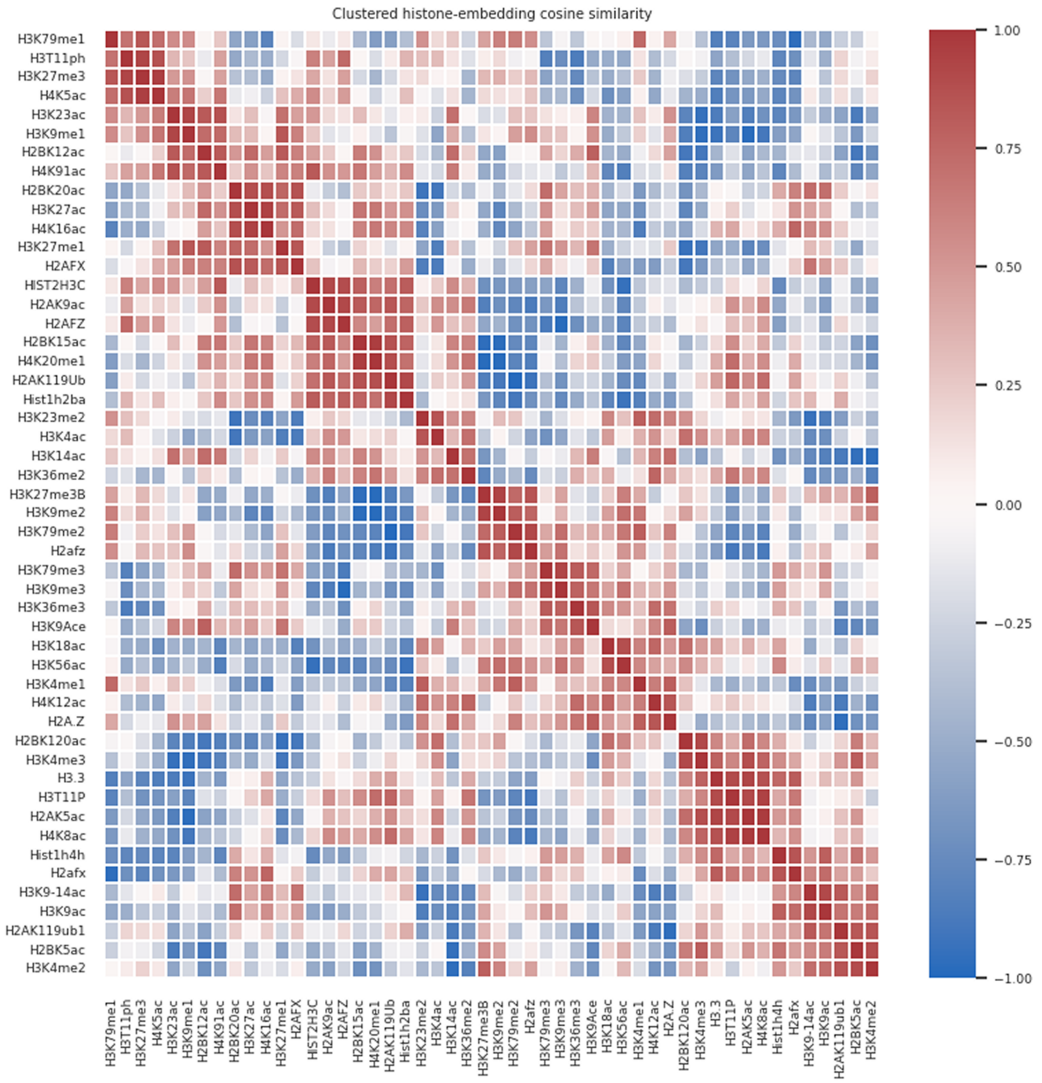


Extended Data Fig. 1 | Distribution of interaction classes across 50 histone modifications. Each bar represents the total number of positive and negative interaction instances for a given histone marker. Positive samples are derived from the RNAInter v4.0 dataset, while negative samples are assumed to be non-interacting.



Extended Data Fig. 2 | Bar chart showing the distribution of RNA categories in the dataset. Each bar (uniform blue) represents the total number of entries for a given RNA type (y-axis), with the exact percentage of the overall dataset printed above each bar. lncRNAs dominate the collection (64.7% of entries), followed by pseudogenes (12.0%), miRNAs (8.9%), and "others" (8.2%); all remaining categories each contribute less than 7% combined.

Each histone modification is encoded as a one-hot categorical vector of length 50. This straightforward encoding allows the model to condition predictions on the specific histone mark under consideration. This baseline representation enables an efficient assessment of whether categorical histone features alone, combined with learned sequence embeddings, are sufficient to capture meaningful patterns of RNA–histone regulation.



Extended Data Fig. 3 | Each cell shows the pairwise cosine similarity between two histone embeddings (n=50), with the diagonal fixed at 1.0. The red regions along the diagonal (mean = +0.0047, s.d. = 0.4001) indicate that **the model has assigned similar embeddings to some histone modifications**, possibly utilizing the information of related histone markers that co-occur on multiple sequences, and as a consequence, the biochemical relationships.

C Training details

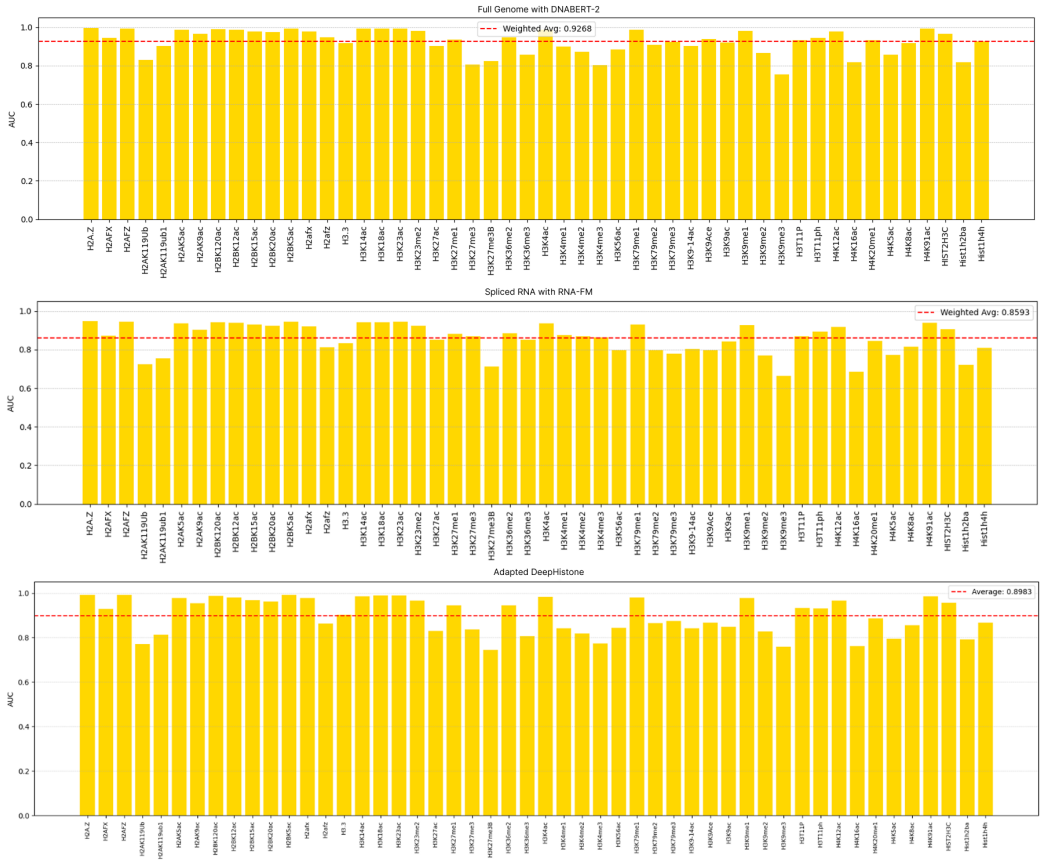
The model is trained to perform binary classification on RNA–histone modification pairs, where the objective is to predict whether a specific histone mark is likely to regulate a given RNA. Each input consists of a concatenated representation of the RNA sequence embedding and a one-hot vector representing the histone modification. These joint vectors are passed through a multi-layer perceptron (MLP) composed of several fully connected layers with ReLU activations, batch normalization, and dropout for regularization Srivastava et al. (2014); Ioffe and Szegedy (2015); Nair and Hinton

(2010). We emphasize implementing the dropout layer after the batch normalization layer, as placing dropout before batch normalization can disrupt the learned activation statistics and reduce training stability. The final layer produces a single logit value, which is transformed using the sigmoid function to yield a probability score. We denote the RNA embedding as $\mathbf{r} \in \mathbb{R}^{d_r}$ and the one-hot histone vector as $\mathbf{h} \in \{0,1\}^{d_h}$. The concatenated input is $\mathbf{x} = [\mathbf{r}; \mathbf{h}] \in \mathbb{R}^{d_r+d_h}$, which is fed into the MLP to produce a prediction $\hat{y} = \sigma(f(\mathbf{x}))$, where f is the MLP and σ is the sigmoid function.

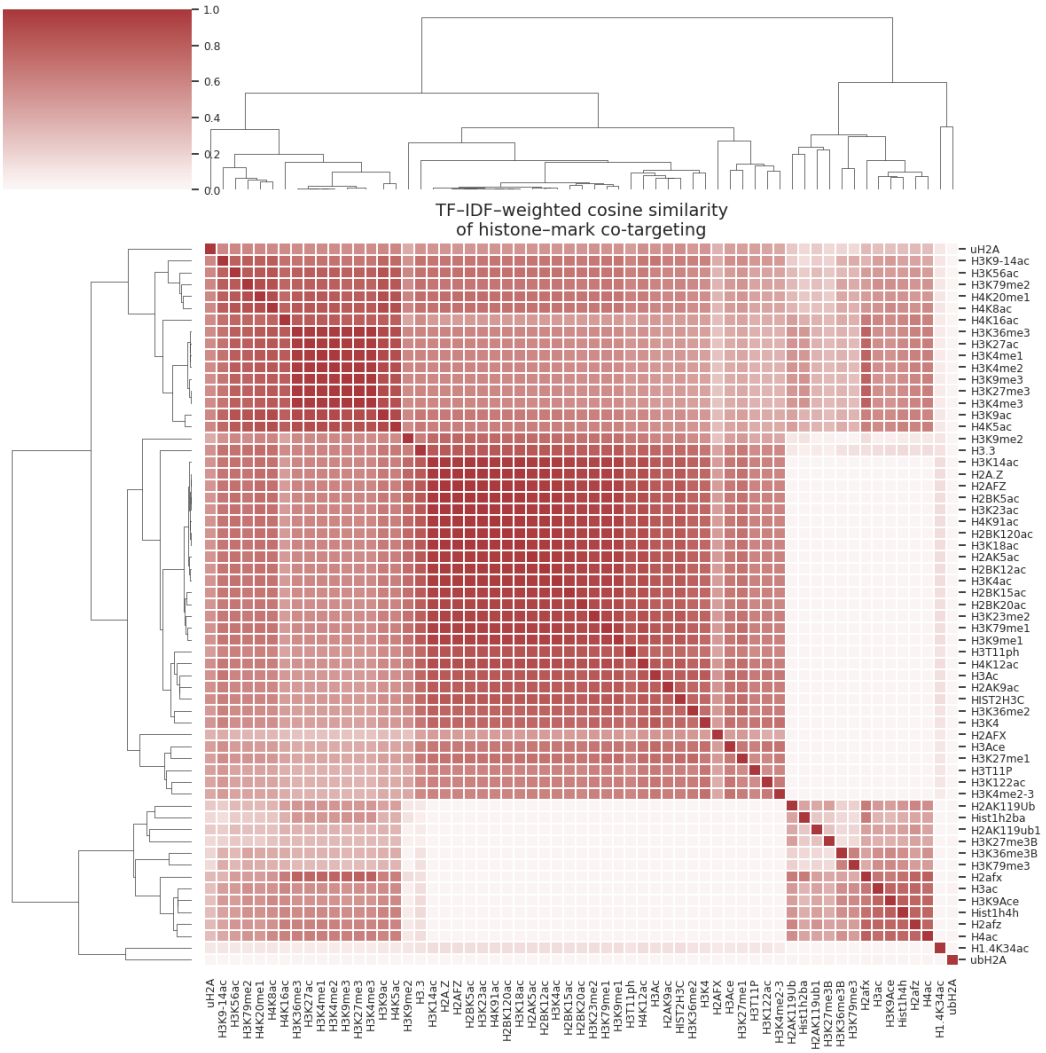
The model is trained using the binary cross-entropy (BCE) loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{1}$$

where $y_i \in \{0,1\}$ denotes the ground truth label for sample i, and $\hat{y}_i$ is the model's predicted probability. We use the AdamW optimizer with a learning rate of 2×10^{-4} and weight decay of 10^{-3}. Training is performed for up to 50 epochs, with early stopping Prechelt (1998), with factor of 0.1 and patience of 3, based on the validation loss to prevent overfitting. Specifically, training is halted if the validation BCE loss does not improve over 5 consecutive epochs. The classifier is run on a single NVIDIA 1080Ti GPU. The embeddings of RNA-FM and DNABERT2 are derived on a single NVIDIA Tesla-V100 GPU. Model selection and early stopping are guided by performance on a validation set held out from the training data. The final evaluation is performed on a disjoint test set of RNA sequences not seen during training or validation, ensuring a strict assessment of generalization to unseen RNAs.



Extended Data Fig. 4 | Per-marker AUROC of DNABERT2 model on gene body + 30 kbp upstream inputs, RNA-FM model on spliced RNA, adapted DeepHistone baseline on 11 kbp promoter + 1 kbp gene body. Macro-averaged AUROC values are also reported.



Extended Data Fig. 5 | TF-IDF-weighted cosine similarity heatmap of histone–mark co-targeting across RNAs. This purely data-driven relation highlights which modifications co-occur in the dataset, with TF–IDF down-weighting ubiquitous marks.

D Baseline implementation and modification

The published DeepHistone source and default hyper-parameters were used without alteration. Inputs comprised 1,000 bp of genomic sequence centred on each TSS, one-hot encoded, without any chromatin-accessibility track. Only one CNN tower was retained (DNA module), the representations were flattened, and passed through a 29-million-parameter dense block with dropout = 0.5 before the final sigmoid. To mitigate over-parameterisation and accommodate a wider regulatory window, we (i) substituted global max- and average-pooling for the flatten operation, concatenating the pooled vectors, and (ii) extended the input field to 11,000 bp upstream of the annotated TSS plus 1,000 bp into the ncRNA gene body. All convolutional filter dimensions and kernel sizes were retained, so the only learnable parameters removed were those in the dense head. The chosen input window size produced the best results. Both vanilla and adapted networks were trained on the RNAInter v4.0 promoter–histone pairs supplied by the original authors. Data splits, loss function, optimiser, and regularisation schedules were kept identical across models; only the architectural change and input length differed. Performance was assessed on the fixed test partition using macro- and micro-averaged AUROC.

E On learning histone representations

Because our model jointly handles all 50 histone modifications as inputs to the model, it offers a unique opportunity to ask whether the network discovers any biochemical relationships among the marks themselves. We inspected a 50, 32, 16, 8, 4 and 2 dimensional trainable histone-marker embeddings after convergence. All model and training configurations were preserved. It is worthwhile to note that the validation loss did not show degradation up to 8-dimensional histone modification embeddings. We went on to construct a heat map of cosine similarity distance of every possible pair of histone modifications and ordered them using hierarchical clustering. A cosine-similarity heat-map (Extended Fig. 3) shows distinctive clusters: average off-diagonal similarity +0.0047 ± 0.4001, with 51.1 % of pairs above |0.25|. Acetylation marks are largely grouped together. Thus, the model has pushed randomly initialized histone representation vectors of histone marks that co-occur (see Fig. 5) to similar vectors. As a part of an ablation experiment, we observed that the histone embeddings remain largely orthogonal if we use under-sampling of the dominant class to balance the training dataset.

F Evaluation Metrics

We evaluate model performance using per-marker and micro-averaged metrics: area under the receiver operating characteristic curve (AUROC), accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC).

In our single-mark framework, we train and evaluate a separate binary classifier for RNA conditioned on each individual histone mark. As a baseline, DeepHistone operates in a multi-label regime, predicting all 50 histone marks simultaneously for each sample. To compare these two settings on equal footing, we use micro-averaged metrics: we pool the predictions and true labels across all marks and samples, and compute global counts of true positives, false positives, false negatives, and true negatives.

Let $C = 50$ denote the number of histone marks, and let $i = 1, \ldots, N$ index samples and $c = 1, \ldots, C$ index classes (marks). For each sample–class pair (i, c), let $y_{i,c} \in \{0, 1\}$ denote the ground-truth label and $\hat{y}_{i,c} \in \{0, 1\}$ the predicted label. We then define the pooled confusion-matrix counts:

$$\mathrm{TP} = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{1}\{y_{i,c} = 1,\ \hat{y}_{i,c} = 1\} \tag{2}$$

$$\mathrm{FP} = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{1}\{y_{i,c} = 0,\ \hat{y}_{i,c} = 1\} \tag{3}$$

$$\mathrm{FN} = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{1}\{y_{i,c} = 1,\ \hat{y}_{i,c} = 0\} \tag{4}$$

$$\mathrm{TN} = \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{1}\{y_{i,c} = 0,\ \hat{y}_{i,c} = 0\} \tag{5}$$

$\mathrm{F1}_{\mathrm{micro}}$ is calculated as a harmonic mean of $\mathrm{Precision}_{\mathrm{micro}}$ and $\mathrm{Recall}_{\mathrm{micro}}$. Varying the threshold value from 0 to 1, we were able to draw a receiver operating characteristic (ROC) curve. The area under this curve was then calculated as a criterion called AUROC. Given the class imbalance present in many histone markers, we highlight MCC and AUROC as key metrics. Unlike accuracy or F1, MCC (Expression 6) accounts for all entries of the confusion matrix and provides a balanced evaluation regardless of class distribution:

$$\frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})} \cdot \sqrt{(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{6}$$

$$\mathrm{AUROC}_{\mathrm{micro}} = \mathrm{AUROC}\big(\{(y_{i,c}, \hat{p}_{i,c})\}_{i=1,c=1}^{N,C}\big), \tag{7}$$

where $\hat{p}_{i,c} \in [0,1]$ is the predicted score for (i,c). To binarize $\hat{p}_{i,c}$, we choose for each class c the threshold that maximizes Youden's J statistic,

$$J = \mathrm{Sensitivity} + \mathrm{Specificity} - 1. \tag{8}$$