ADAREDIT: A Graph Foundation Model for Interpretable A-to-I RNA Editing Prediction

Zohar Rosenwasser
Bar-Ilan University
Israel
rosenwz@biu.ac.il

Erez Y. Levanon
Bar-Ilan University
Israel
erez.levanon@biu.ac.il

Michael Levitt
Stanford University
United States
levittm@stanford.edu

Gal Oren
Stanford University, Technion
United States
galoren@stanford.edu

Abstract

Adenosine-to-inosine (A-to-I) RNA editing, catalyzed by ADAR enzymes, is a prevalent post-transcriptional modification with roles in transcript stability, splicing, and protein recoding. Accurate prediction of editing sites remains difficult due to the intricate interplay between local sequence context and RNA secondary structure. Existing approaches either rely on brittle, handcrafted features or adapt generic foundation models trained on broad RNA datasets, which fail to capture the specific biochemical requirement for double-stranded RNA and often lack interpretability.

We introduce ADAREDIT, a domain-specialized graph foundation model for A-to-I editing site prediction. RNA segments are represented as graphs with nucleotides as nodes and both sequential and base-pairing edges, enabling the model to learn biologically aligned features such as stem–loop motifs. A Graph Attention Network architecture yields mechanistic interpretability by highlighting influential structural and sequence neighbors. Across 25 cross-tissue evaluations, ADAREDIT consistently outperforms prior methods ($F_1 > 0.85$) and generalizes to evolutionarily distant species, demonstrating that biology-aware foundation models can deliver superior accuracy, scalability, and insight for complex RNA modification tasks. The sources of this work are available at our repository: https://github.com/Scientific-Computing-Lab/AdarEdit.

1 Introduction

Adenosine-to-inosine (A-to-I) RNA editing is a widespread post-transcriptional mechanism that modifies pre-mRNA molecules encoded in the genome. Such editing events can alter amino acid sequences, affect alternative splicing, regulate gene silencing, and influence RNA stability and localization [1, 2]. In animals, A-to-I editing is catalyzed by the adenosine deaminase acting on RNA (ADAR) family of enzymes [3, 4, 5, 6, 7], which bind double-stranded RNA (dsRNA) structures and deaminate specific adenosines (A) to inosines (I), interpreted during translation as guanosines (G). Dysregulated A-to-I editing has been implicated in cancer [8, 9, 10, 11, 12], neurological disorders [13, 14, 15], and autoimmune conditions. ADAR proteins contain two key domains: a double-stranded RNA-binding domain (dsRBD), responsible for recognizing dsRNA, and a deaminase domain, which catalyzes the hydrolytic deamination reaction [3, 16].

Mammals encode three ADAR isoforms: ADAR1 (ubiquitous; long-dsRNA substrates), ADAR2 (many site-specific recoding events, e.g., in *GRIA2* and *FLNA*), and ADAR3 (catalytically inactive, potentially inhibitory) [3, 16]. Distinct expression patterns across tissues produce characteristic editing landscapes. In humans, the majority of A-to-I editing occurs within *Alu* repetitive elements,

Submitted to Multi-modal Foundation Models and Large Language Models for Life Sciences at NeurIPS 2025.

which make up roughly 10% of the genome. When *Alu* elements are present in inverted orientations within the same transcript, they can form stable dsRNA structures through intramolecular base pairing, creating ideal ADAR substrates [6, 5]. Editing specificity and efficiency are shaped by local sequence motifs, nearest-neighbor preferences, and structural features such as bulges or mismatches that disrupt perfect base-pairing [8, 17]. Yet, the precise determinants guiding ADAR enzymes to target specific adenosines remain incompletely understood [4, 18].

Programmable RNA editing leverages endogenous ADAR via antisense guide RNAs that hybridize to target transcripts, creating dsRNA and enabling site-specific A-to-I conversion. Because edits occur at the RNA rather than the DNA level, they are reversible and avoid permanent genomic change, potentially reducing safety risks and off-target effects. Success hinges on accurately predicting ADAR-compatible substrates, requiring computational models that integrate sequence and structural determinants to guide gRNA design.

Recent advances in foundation models have shown remarkable success across diverse RNA tasks (**§Appendix A**). Meanwhile, editing-site prediction has mainly followed two routes — engineered-feature classifiers and large pre-trained sequence/RNA language models usage or fine-tuning. The former are brittle across biological contexts; the latter treat RNA as a linear string and only indirectly encode structure, limiting ADAR-specific accuracy and interpretability (**§Appendix B**).

Graph neural networks (GNNs) [19, 20, 21, 22] provide a natural architecture for this task by representing RNA secondary structures as graphs, with nucleotides as nodes and both sequential and base-pairing interactions as edges. This encoding reflects the biochemical reality of ADAR recognition, enabling models to capture relationships between editing sites, base-pairing partners, loop regions, and surrounding sequence context.

Here, we present ADAREDIT, a Graph Attention Network (GAT)-based model [23] for A-to-I editing site prediction. Unlike generic foundation models, ADAREDIT is *pretrained across a broad spectrum of editing-specific datasets* covering multiple human tissues with distinct ADAR isoform profiles, as well as diverse non-human species. This broad, biologically grounded pretraining yields a single model that *generalizes without retraining* to unseen tissues, conditions, and species, satisfying the defining criterion of a foundation model in computational biology. Its graph attention mechanism enables scalable learning of editing determinants while providing interpretable insights into both established and novel structural motifs. Once trained, ADAREDIT can be directly adapted to new RNA editing prediction tasks, guide RNA design, or other double-stranded RNA-related analyses *without task-specific re-engineering*, offering a reusable foundation for the RNA editing domain. By uniting sequence and structure in a biologically faithful representation, ADAREDIT advances predictive performance, interpretability, and transferability — hallmarks of a true domain-specific foundation model.

2 Model Structure

Graph-Based RNA Representation. ADAREDIT employs a novel graph-based approach to represent RNA editing contexts by encoding RNA segments as structured networks that capture both sequential and structural relationships. Each RNA segment is represented as a graph where individual nucleotides serve as nodes, connected by two distinct types of edges: sequential edges linking adjacent nucleotides in the 5' to 3' direction (creating bidirectional connections between consecutive bases), and structural edges connecting base-paired nucleotides as predicted by computational RNA folding algorithms using dot-bracket notation parsing. This dual-edge architecture enables the model to simultaneously consider local sequence context and long-range structural interactions that are critical for ADAR enzyme recognition and binding. The GAT architecture consists of three stacked GAT layers, each employing multi-head attention mechanisms (4 attention heads per layer) to learn hierarchical and context-aware feature representations. Each node (nucleotide) is initially encoded with an 8-dimensional feature vector comprising: one-hot encoding for the four RNA bases plus unknown nucleotides (N) totaling 5 dimensions, binary pairing status indicating whether the nucleotide participates in base-pairing (1), relative positional distance from the candidate editing site (1), and a binary target site flag identifying the candidate adenosine (1).

Through the three GAT layers, each node iteratively aggregates information from its graph neighbors via learned attention weights, with each layer applying ReLU activation, batch normalization, and dropout (rate=0.2) for regularization. This multi-layer architecture enables the model to learn hierarchical feature representations, where each successive layer can potentially integrate information from an expanding neighborhood within the graph structure. The attention mechanism dynamically

weights the importance of different sequence and structural contexts, enabling the model to learn which nucleotide relationships are most predictive for A-to-I editing events.

The final graph-level representation is obtained through global mean pooling across all node embeddings, which aggregates the learned features from the entire RNA segment into a fixed-size vector. This pooled representation is then passed through a fully connected layer that maps the high-dimensional graph features to a single output dimension, followed by a sigmoid activation function to produce the binary classification probability for editing prediction. The model also extracts attention weights from the first GAT layer during inference, providing interpretable insights into which nucleotide relationships contribute most strongly to each editing prediction (Illustrative diagram is provided in **§Appendix C**-Figure 1.)

3 Datasets, Evaluations and Results

Datasets. We used two distinct dataset types:
1. Human tissue-specific Alu datasets. We curated a high-confidence catalogue of double-stranded *Alu* substrates, following the protocol in previous work [24], focusing on *Alu* pairs whose predictable secondary structures reliably identify optimal ADAR substrates with well-defined base-pairing. The full human dataset construction process is provided in **§Appendix D**, and summary counts and ADAR isoform profiles are provided in Table 1.
2. Cross-species non-Alu datasets. To test generalizability beyond the human *Alu*-based context, we constructed datasets for three evolutionarily distant species lacking *Alu* elements. These taxa represent widely separated phylogenetic lineages (echinoderms, hemichordates, mollusks) and have been extensively documented to harbor numerous A-to-I editing sites [25], providing a stringent test of cross-species generalization. In these species, editing often occurs within non-*Alu* repetitive sequences or other double-stranded RNA structures. The complete cross-species dataset construction pipeline is described in **§Appendix E** and summary statistics are provided in Table 2.

Evaluations and Results.
1. Cross-tissue evaluation (human data). To evaluate the capacity of ADAREDIT to learn and generalize editing patterns across diverse biological contexts, we constructed five separate tissue-specific subsets: Brain Cerebellum, Artery Tibial, Liver, Muscle Skeletal, and a Combined dataset integrating editing data from all 47 GTEx tissues. We trained ADAREDIT independently on each of the five tissue-specific datasets and conducted cross-validation evaluations, testing each tissue-specific model against each of the four other tissues and the Combined dataset — resulting in 25 train–validation experiments. This design enabled systematic assessment of the model's generalization capability across tissues with distinct ADAR compositions, as well as evaluation of its behavior when trained on editing profiles from diverse contexts. For all experiments, model selection was performed per epoch on the validation set using a composite *Performance Score* (see **§Appendix F**).

The experimental results are presented in **§Appendix G**-Figure 2-A. Generally, the highest performance was observed when ADAREDIT was trained and evaluated on data derived from the same tissue. This result is expected, as ADAR expression profiles and editing patterns vary among tissues, reflecting biological features specifically tailored to each tissue's context. The highest overall performance (ACC=0.86, REC=0.85, F1=0.85, and PRE=0.86) was achieved when the model was both trained and evaluated on the combined dataset, integrating data from all tissues.

Furthermore, the consistently lowest predictive performance (F1 = 0.76–0.80) was observed for the Muscle Skeletal dataset. This reduced performance is most likely due to the smaller dataset size available for this tissue compared to the larger datasets used for the other tissues. As is common in foundation models, ADAREDIT exhibits a strong dependence on training data volume — performance scales with both the quantity and diversity of data — suggesting that substantially more comprehensive coverage for this tissue would lead to marked improvements.

Another notable trend was the reduction in model accuracy when training and validation were performed between tissues with distinct ADAR isoform expression profiles. For example, a model trained on Liver tissue showed reduced performance when evaluated on Artery Tibial, and vice versa. This cross-tissue drop in performance (accuracy 0.80 versus 0.83 in within-tissue settings) highlights the model's sensitivity to isoform-specific editing patterns, indicating that it captures biologically relevant features linked to the tissue-specific activity of ADAR isoforms.
2. Cross-species evaluation (non-human data). Within-species training and validation yielded high predictive performance: *S. purpuratus* (F1 = 0.85), *P. flava* (F1 = 0.84), and *O. bimaculoides* (F1 = 0.87) (**§Appendix G**-Figure 2-B), comparable to the highest scores obtained in human tissue-specific models. Model selection was also applied here as well, as detailed in **§Appendix F**. These

results demonstrate that ADAREDIT's graph-based representation and attention mechanisms capture biologically meaningful RNA editing patterns even in the absence of *Alu* elements.

We next trained ADAREDIT exclusively on the human Combined *Alu* dataset and evaluated it on each non-human species. As expected, cross-species performance declined relative to within-species results, with F1 scores dropping by ∼0.11–0.17 to values between 0.68 and 0.76 (**§Appendix G**-Figure 2-B). Nonetheless, the human-trained model retained moderate predictive power, indicating partial conservation of fundamental sequence–structure signals recognized by ADAR enzymes across deep evolutionary distances. Overall, these cross-species experiments reveal that while RNA editing patterns are partly conserved, they also exhibit strong species-specific components. This duality highlights the importance of incorporating diverse species data to build models with both broad generalization capability and fine-grained adaptation to organism-specific editing contexts.

4 Interpretability of ADAREDIT via Attention Analysis

A central motivation for employing an attention-based architecture in ADAREDIT was to couple high predictive accuracy with the ability to extract biologically interpretable signals. Unlike black-box models, attention mechanisms can directly highlight the sequence–structure relationships most relevant to the model's decisions [26, 27], offering a window into the molecular features that govern ADAR specificity. In ADAREDIT, each edge between nucleotides is assigned an attention coefficient α_{ij}, quantifying the influence of neighbor node j on the updated representation of target node i.

Given the ongoing debate on whether attention weights directly constitute explanations [28, 29], it is not trivial to assume that learned attention coefficients automatically reflect biologically meaningful determinants. To explicitly assess this in ADAREDIT (**§Appendix H** Figure 3-A), we first extracted attention coefficients from the initial Graph Attention layer. For each nucleotide position within a 1,200-nt window centered on the candidate editing site (from –600 to +600), we computed the maximal attention weight across all edges originating from that position. This produced a positional attention profile indicating the relative importance of each location in predicting editing status.

We next trained an independent XGBoost classifier [30] directly using these positional attention features. Although performance slightly declined relative to the full ADAREDIT model, the classifier trained solely on attention-derived features still achieved robust predictive accuracy. For instance, using the combined Alu dataset, the attention-based XGBoost model yielded an F1-score of 0.81, accuracy of 0.8, precision of 0.76, and recall of 0.87 (**§Appendix H** Figure 3-B).

To gain deeper interpretability, we then applied SHapley Additive exPlanations (SHAP) analysis [31] – a method that quantifies the contribution of each feature to the model's predictions – to identify the most influential attention features (**§Appendix H** Figure 3-C). SHAP analysis revealed that positions immediately surrounding the edited adenosine (positions 0, +1, and 1) consistently ranked among the top features, aligning with previous experimental findings identifying these nucleotides as critical determinants of editing efficiency. Intriguingly, additional distal positions also emerged as influential, highlighting previously underappreciated structural motifs that may warrant further biological investigation.

Finally, we retrained the XGBoost classifier using only the top 20 SHAP-identified attention features. Despite the substantial reduction in input features, model performance remained high with only a modest decline (F1-score 0.82, accuracy 0.81, precision 0.79, and recall 0.85. The ability of such a small subset of attention-derived features to approximate full model accuracy strongly suggests that ADAREDIT inherently identifies and focuses upon a biologically meaningful subset of sequence-structure relationships. In addition to this analysis, we analyzed the positional distribution of attention weights around each editing site as described in **§Appendix I**.

Acknowledgments

This work was supported by US National Institutes of Health award R35GM122543 (M.L.), Foundation Fighting Blindness (grants TA-GT-0620-0790-HUJ and PPA-0923-0865-HUJ), grants from the Israeli Ministry of Science (grant 3-17916), and by Israel Science Foundation (2637/2) (E.Y.L.). M.L. is the Robert W. and Vivian K. Cahill Professor of Cancer Research. E.Y.L. is a fellow at the Israel Institute of Advanced Studies.

A Appendix: Previous Work — Foundation Models for RNA Tasks

RNA can be conceptualized as a structured biological language, where primary sequences encode functional and regulatory information and higher-order structures mediate biological activity. This duality makes RNA a natural target for foundation models (FMs) that learn general-purpose representations from large unlabeled corpora, and for generative models that can design novel, function-bearing sequences under biochemical constraints. Inspired by advances in natural language processing, recent work has explored both encoder-based and generative architectures tailored to RNA data.

Encoder-only RNA foundation models leverage masked language modeling or contrastive objectives to capture sequence–structure dependencies. RNABERT [32] learns embeddings enriched with structural information, improving family classification and structure-aware tasks. RNAErnie [33] integrates motif-aware pretraining, enabling transfer to multiple RNA-related problems. RNA-FM [34], trained on over 23 million sequences, has been integrated into downstream structure predictors such as RhoFold+ [34] to improve RNA 3D modeling. Other specialized large-scale models include RiNALMo [35], a 650M-parameter model that generalizes across secondary-structure and binding tasks.

Generative and instruction-tuned RNA models focus on design, optimization, and interactive analysis. GenerRNA [36] is a Transformer-based generator for de novo RNA design with controllable structural features. RNA-GPT [37] aligns RNA encoders with general-purpose large language models to support multimodal RNA question-answering and guided sequence editing. RiboDiffusion [38] conditions diffusion models on RNA 3D backbones for inverse folding.

These models have been applied in diverse RNA contexts. In structure prediction, RhoFold+ [34] integrates RNA-FM embeddings into a transformer–invariant point attention pipeline, achieving state-of-the-art performance on RNA-Puzzles and CASP targets. For RNA–protein interaction modeling, BERT-RBP [39] predicts RBP binding sites from sequence while revealing attention patterns linked to structural motifs. Generative fine-tuning has been applied to propose mutations in rRNA with RNA language models [40], while mRNA optimization frameworks such as GEMORNA [41] and RiboCode [42] improve translation efficiency and vaccine design.

Most general RNA FMs treat RNA as a linear string and incorporate structure only indirectly; design-oriented generators often optimize objectives not fully aligned with the biochemical determinants of ADAR-mediated editing (**§Appendix B** GPT usage for the problem, for example). These gaps motivate the development of domain-specific architectures – such as graph-based FMs trained on editing-labeled data – that retain the scalability and transferability of foundation models while encoding mechanistic, biology-aware inductive biases.

B Appendix: Previous Work — A-to-I RNA Editing Site Prediction

Computational prediction of A-to-I editing sites has evolved through several distinct methodological phases. Early approaches relied on classical machine learning algorithms such as support vector machines, random forests, and XGBoost, utilizing handcrafted features including sequence motifs, nucleotide composition around editing sites, thermodynamic properties of RNA secondary structures, and evolutionary conservation scores [43, 44, 45, 46]. While these feature-based methods achieved moderate predictive performance, their dependence on predefined biological features limited their ability to discover novel editing determinants and required extensive manual re-engineering when applied across different biological contexts or species.

The advent of deep learning introduced convolutional and recurrent neural networks that attempted to automatically learn sequence patterns without explicit feature engineering [47, 44]. However, these early deep learning approaches still fundamentally treated RNA as a linear sequence, failing to adequately incorporate the three-dimensional structural context critical for ADAR enzyme recognition and substrate binding.

More recent efforts have adapted large pre-trained language models specifically for A-to-I editing site prediction. GPT-based models have been applied to this task by framing editing prediction as a sequence classification problem [24, 48]. In our evaluation using the same dataset employed in this study, GPT-4o mini achieved an accuracy of 69.9%, precision of 65.4%, recall of 81.6%, and an F1-score of 72.7%, demonstrating substantial pattern recognition capabilities while revealing significant room for improvement.

Despite their promise, current foundation model approaches face critical limitations when applied to RNA editing prediction. Most models are trained on generic RNA sequences with objectives unrelated to A-to-I editing biochemistry, and even when secondary structure information is provided as additional input, they fundamentally represent RNA as a linear sequence rather than capturing

the spatial relationships crucial for ADAR recognition. Furthermore, fine-tuned foundation models offer limited interpretability regarding the biological mechanisms driving their predictions, making it difficult to validate learned representations against established ADAR biology. These limitations highlight the need for computational frameworks that natively integrate RNA sequence and structure information while being trained specifically on editing-relevant data to capture the unique biochemical constraints of ADAR-mediated A-to-I editing.

C Appendix: Model Architecture and Graph Construction

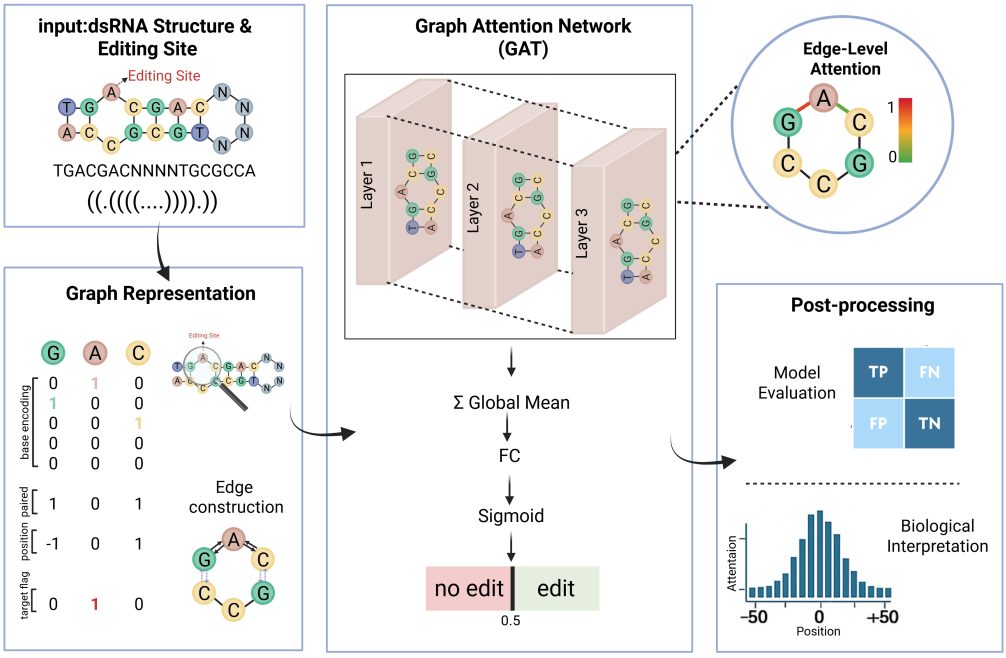

Figure 1: **ADAREDIT Model Architecture and RNA Graph Construction.** The figure illustrates the complete ADAREDIT workflow for RNA editing prediction. Input (top left): dsRNA structure with candidate editing site (red 'A') and corresponding dot-bracket secondary structure notation. Graph Representation (bottom left): Conversion of RNA sequence into graph format where nucleotides are nodes connected by edges representing sequential (solid lines) and structural base-pairing (dashed lines) relationships. Each node contains an 8-dimensional feature vector including base encoding, pairing status, relative position, and target flag. Graph Attention Network (center): Three-layer GAT architecture with multi-head attention (4 heads per layer) processing the input graph through successive layers (Layer 1, 2, 3), followed by global mean pooling (Σ Global Mean) and fully connected layer (FC) with sigmoid activation. Edge-Level Attention (circular detail): Visualization of learned attention weights between connected nucleotides, with color intensity indicating attention strength (scale 0-1). Post-processing (right): Model evaluation metrics (confusion matrix: TP, TN, FP, FN) and biological interpretation showing attention distribution across nucleotide positions relative to the editing site (-50 to +50), enabling identification of sequence motifs important for A-to-I editing prediction.

D Appendix: Human Dataset Construction (Alu)

To construct the dataset, we systematically scanned all human UTR regions to identify the closest pair of oppositely oriented Alu elements within each UTR, maximizing the likelihood of stable dsRNA formation. This stringent selection process yielded a total of 905 Alu pairs. For each selected pair, we predicted the secondary structure using RNAfold [49], obtaining a clearly defined duplex structure that served as the structural input for ADAREDIT graph representation.

Following structure prediction, we extracted editing levels for each adenosine within these duplexes from the GTEx RNA-seq dataset (8,603 RNA-seq samples across 47 tissues from 548

donors) [50], both for each tissue separately and for a combined set integrating all tissues. The editing level for each adenosine was calculated as the ratio of reads identifying the adenosine as guanosine (G) relative to the total number of reads covering that position (A+G). Subsequently, we retained for analysis only those sites supported by at least 100 sequencing reads. The final curated dataset comprised 127,015 adenosines. Adenosines with editing levels $\geq 10\%$ were defined as edited, while those with editing levels $< 1\%$ were defined as unedited following thresholds established in previous studies [24].

Table 1: Human tissue datasets for training and evaluation of ADAREDIT; train/validation split is 8/2.

Tissue	**ADAR isoform profile**	**Total sites**	**Edited:Unedited**	**Train:Val**
Brain Cerebellum	ADAR1+ADAR3 high	24,000	12,000:12,000	19,200:4,800
Artery Tibial	ADAR2-dominant	24,000	12,000:12,000	19,200:4,800
Liver	ADAR1 high	24,000	12,000:12,000	19,200:4,800
Muscle Skeletal	ADAR1 low	7,250	3,625:3,625	5,800:1,450
Combined	Mixed	24,000	12,000:12,000	19,200:4,800

E Appendix: Cross-species Dataset Construction

For each organism, we constructed an annotated dataset using previously reported editing sites and their measured editing levels from Zhang *et al.* [25]. First, we merged proximal editing sites within 1,000 bp intervals to form discrete editing clusters. For each cluster, we extracted an extended sequence including an additional 1,000 bases on each side and predicted the minimum free-energy secondary structure using RNAfold. We then selected the segment containing the highest density of editing sites within each folded structure and extracted all adenosines that met our stringent editing criteria (editing level $\geq 10\%$ and sequencing coverage $\geq 100\%$ reads). Non-edited adenosines located within 20 bases of an edited site were labeled as negative examples. The resulting datasets were structurally and format-wise analogous to the human Alu dataset, enabling direct comparison.

Table 2: Cross-species datasets for evaluating generalization; train/validation split is 8/2.

Species	**Total sites**	**Edited:Unedited**	**Train:Val**
Strongylocentrotus purpuratus	15086	7543:7543	12068:3018
Ptychodera flava	9112	4556:4556	7289:1823
Octopus bimaculoides	39472	19736:19736	31577:7895

F Appendix: Model Selection

At the end of each training epoch we evaluated the current model on the validation set and recorded Accuracy, F1, Sensitivity (Recall), Specificity, and Precision, using a 0.5 threshold to binarize predictions. For model selection, checkpoints were saved at fixed intervals of 10 epochs, and for each saved checkpoint we computed a composite *Performance Score* defined as the unweighted sum of the five validation metrics, $\mathrm{PerformanceScore}_e = \mathrm{Accuracy}_e + \mathrm{F1}_e + \mathrm{Sensitivity}_e + \mathrm{Specificity}_e + \mathrm{Precision}_e$. The checkpoint with the highest score among these 10-epoch snapshots was selected for reporting (in case of a tie, the earliest checkpoint was preferred). The selected checkpoint was then used consistently for all evaluations and analyses, including the results shown in Figure 2 and the attention-based interpretability in Figure 3.

G Appendix : Cross-Tissue and Cross-Species Performance

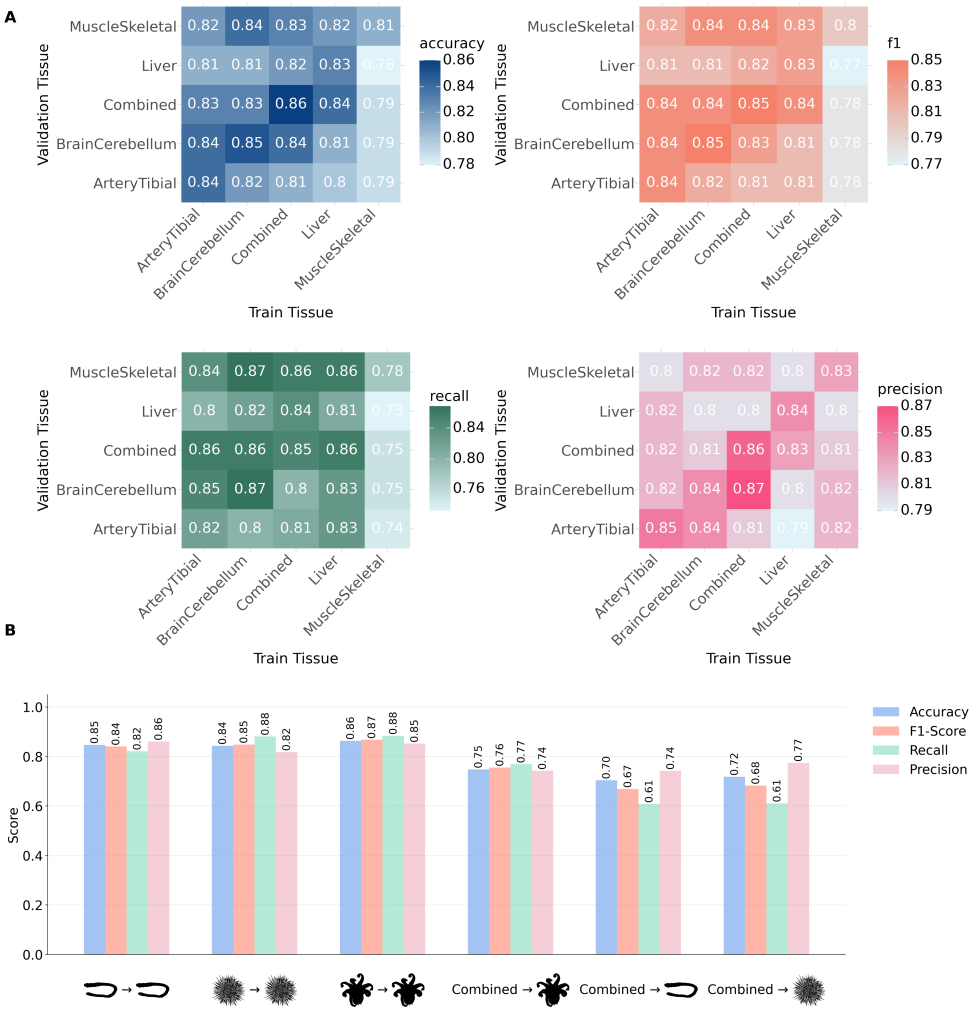

Figure 2: **Cross-tissue and cross-species evaluation of ADAREDIT.** (A), Cross-tissue performance of ADAREDIT when trained and validated on each pair of tissue-specific datasets. Four heatmaps present accuracy, F1-score, recall, and precision for all train-validation tissue combinations: Muscle Skeletal, Liver, Combined (all GTEx tissues), Brain Cerebellum, and Artery Tibial. Highest values are generally observed for within-tissue evaluations, with the Combined dataset achieving the top overall scores. (B), Cross-species evaluation of ADAREDIT. Three left bar groups: within-species performance for Strongylocentrotus purpuratus, Ptychodera flava, and Octopus bimaculoides when trained and validated on the same species dataset. Three right bar groups: cross-species performance when trained on the human Combined Alu dataset and validated on each non-human species. Bars represent accuracy, F1-score, recall, and precision.

H Appendix: Attention-Based Model Interpretability

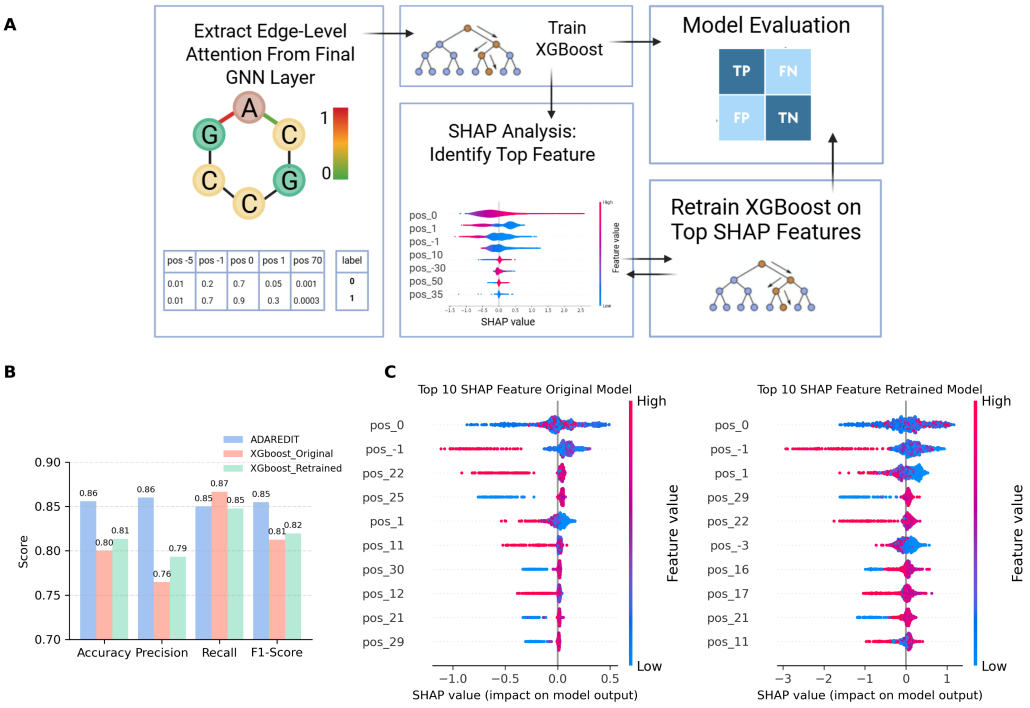

Figure 3: **Interpretability analysis of ADAREDIT attentions.** (A) Workflow for extracting edge-level attention coefficients from the GAT, aggregating them into positional attention features, training an XGBoost classifier, performing SHAP analysis to identify top features, and retraining the classifier using only these features. (B) Comparison of predictive performance (accuracy, precision, recall, F1-score) for ADAREDIT, the original attention-based XGBoost model, and the retrained model using top SHAP features, showing that ADAREDIT achieves the highest scores while the retrained XGBoost maintains near-identical performance to the full-feature model, indicating that a small subset of attention-derived features retains predictive power. (C) SHAP summary plots for the top 10 attention-derived features in the original XGBoost model (left) and in the retrained model (right), showing feature importance and direction of influence. Results shown are for the Combined→Combined dataset in human Alu sequences.

I Appendix : Positional Attention Analysis

Having established that attention mechanisms embedded within the ADAREDIT model can provide meaningful explanations for editing site prediction, we conducted a deeper analysis of attention profiles to identify known and potentially novel biological determinants influencing ADAR-mediated editing. Using the attention scores extracted previously for positions spanning 50 nucleotides upstream and downstream of each editing site (position 0), we performed several targeted analyses on the attention data derived from the Combined Alu model, which showed the highest predictive performance (Figure 4).

First, we examined mean positional attention across all sites (Figure 4-a). As anticipated, position 0 – the edited adenosine itself – exhibited significantly higher average attention compared to surrounding nucleotides. This result aligns with existing biological knowledge indicating that structural properties of the editing site (particularly its presence in a loop or an unpaired region) strongly influence editing efficiency. Interestingly, we also observed slightly elevated attention scores immediately downstream of the editing site, suggesting additional positional influences that merit future exploration.

Next, we analyzed attention scores according to nucleotide identity (Figure 4-b). We found consistently higher average attention at positions containing G or C bases compared to those con-

taining A or T, possibly reflecting the greater stability conferred by GC pairing in dsRNA structures. Particularly noteworthy is the elevated attention observed at position -1 when occupied by guanosine (G), consistent with previous experimental findings demonstrating that a G immediately upstream of the editing site can dramatically reduce or abolish editing.

Further investigation into structural context (paired vs. loop/unpaired positions) showed clear differences in attention distribution (Figure 4-c). Loop or unpaired positions consistently received higher mean attention scores throughout the examined window compared to positions engaged in base pairing. This observation reinforces the biological significance of local RNA structure as a major determinant of editing efficiency.

Additionally, when comparing mean attention profiles between edited and unedited sites (Figure 4-d), we detected no substantial differences.

Lastly, we evaluated attention scores specifically within loop regions, categorized by loop size (Figure 4-e). Positions within loops larger than one nucleotide showed progressively increased mean attention scores, with loops containing two or more nucleotides receiving notably higher attention. This result suggests that ADAREDIT recognizes larger loop structures as particularly informative regions, emphasizing their biological relevance in guiding ADAR activity.

Together, these analyses confirm that attention weights derived from ADAREDIT provide biologically meaningful insights, highlighting previously known determinants of RNA editing efficiency and identifying novel structural and sequence motifs worthy of future experimental exploration.

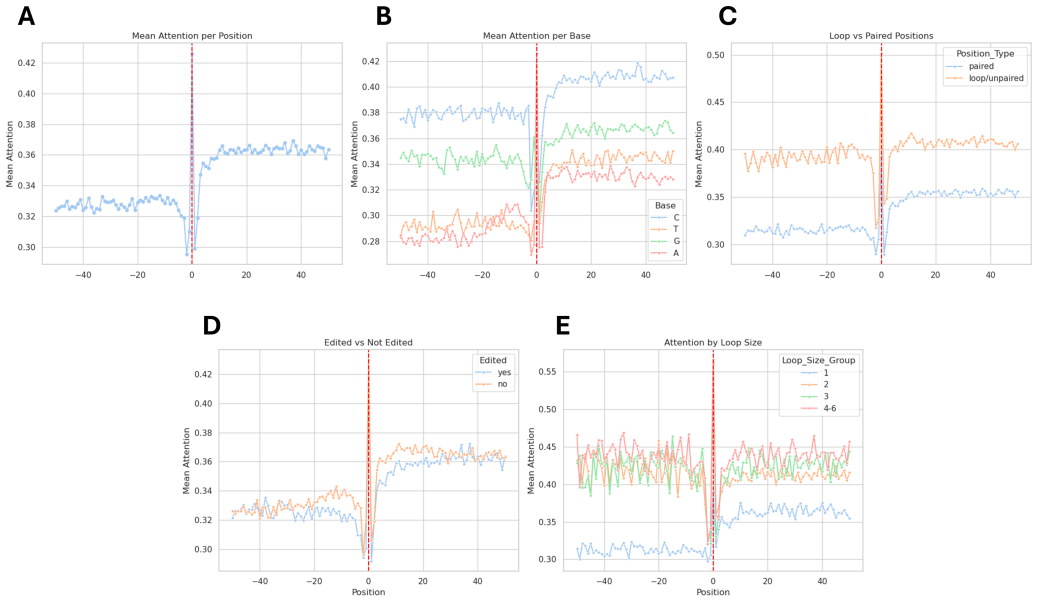

Figure 4: **Positional attention analysis around RNA editing sites.** All graphs depict mean attention scores derived from the ADAREDIT model, spanning 50 nucleotides upstream and downstream of the editing site (position 0): (A) Overall mean attention at each position relative to the editing site. (B) Mean attention scores stratified by nucleotide identity (A, T, C, G). (C) Mean attention scores comparing loop/unpaired versus base-paired positions. (D) Mean attention comparison between edited and unedited sites. (E) Mean attention scores for positions located within loops/unpaired regions, categorized by loop size (1, 2, 3, and 4–6 nucleotides).

References

[1] Axel Brennicke, Anita Marchfelder, and Stefan Binder. Rna editing. *FEMS microbiology reviews*, 23(3):297–316, 1999.

[2] Jonatha M Gott and Ronald B Emeson. Functions and mechanisms of rna editing. *Annual review of genetics*, 34(1):499–531, 2000.

[3] Brenda L Bass. Rna editing by adenosine deaminases that act on rna. *Annual review of biochemistry*, 71(1):817–846, 2002.

[4] Eli Eisenberg and Erez Y Levanon. A-to-i rna editing—immune protector and transcriptome diversifier. *Nature Reviews Genetics*, 19(8):473–490, 2018.

[5] Lily Bazak, Erez Y Levanon, and Eli Eisenberg. Genome-wide analysis of alu editability. *Nucleic acids research*, 42(11):6876–6884, 2014.

[6] Erez Y Levanon, Eli Eisenberg, Rodrigo Yelin, Sergey Nemzer, Martina Hallegger, Ronen Shemesh, Zipora Y Fligelman, Avi Shoshan, Sarah R Pollock, Dan Sztybel, et al. Systematic identification of abundant a-to-i editing sites in the human transcriptome. *Nature biotechnology*, 22(8):1001–1005, 2004.

[7] Hagit T Porath, Shai Carmi, and Erez Y Levanon. A genome-wide map of hyper-edited rna reveals numerous new sites. *Nature communications*, 5(1):4726, 2014.

[8] Julie M Eggington, Tom Greene, and Brenda L Bass. Predicting sites of adar editing in double-stranded rna. *Nature communications*, 2(1):319, 2011.

[9] Debora Fumagalli, David Gacquer, Françoise Rothé, Anne Lefort, Frederick Libert, David Brown, Naima Kheddoumi, Adam Shlien, Tomasz Konopka, Roberto Salgado, et al. Principles governing a-to-i rna editing in the breast cancer transcriptome. *Cell reports*, 13(2):277–289, 2015.

[10] Leng Han, Lixia Diao, Shuangxing Yu, Xiaoyan Xu, Jie Li, Rui Zhang, Yang Yang, Henrica MJ Werner, A Karina Eterovic, Yuan Yuan, et al. The genomic landscape and clinical relevance of a-to-i rna editing in human cancers. *Cancer cell*, 28(4):515–528, 2015.

[11] Domenico Alessandro Silvestris, Ernesto Picardi, Valeriana Cesarini, Bruno Fosso, Nicolò Mangraviti, Luca Massimi, Maurizio Martini, Graziano Pesole, Franco Locatelli, and Angela Gallo. Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma. *Genome biology*, 20:1–18, 2019.

[12] William Slotkin and Kazuko Nishikura. Adenosine-to-inosine rna editing and human disease. *Genome medicine*, 5:1–13, 2013.

[13] Prashant Kumar Srivastava, Marta Bagnati, Andree Delahaye-Duriez, Jeong-Hun Ko, Maxime Rotival, Sarah R Langley, Kirill Shkura, Manuela Mazzuferi, Bénédicte Danis, Jonathan van Eyll, et al. Genome-wide analysis of differential rna editing in epilepsy. *Genome research*, 27(3):440–450, 2017.

[14] Yukio Kawahara, Kyoko Ito, Hui Sun, Hitoshi Aizawa, Ichiro Kanazawa, and Shin Kwak. Rna editing and death of motor neurons. *Nature*, 427(6977):801–801, 2004.

[15] Gilad Silberberg, Daniel Lundin, Ruth Navon, and Marie Öhman. Deregulation of the a-to-i rna editing mechanism in psychiatric disorders. *Human molecular genetics*, 21(2):311–321, 2012.

[16] Kazuko Nishikura. Functions and regulation of rna editing by adar deaminases. *Annual review of biochemistry*, 79(1):321–349, 2010.

[17] Katrina A Lehmann and Brenda L Bass. Double-stranded rna adenosine deaminases adar1 and adar2 have overlapping specificities. *Biochemistry*, 39(42):12875–12884, 2000.

[18] Kazuko Nishikura. A-to-i editing of coding and non-coding rnas by adars. *Nature reviews Molecular cell biology*, 17(2):83–96, 2016.

[19] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[20] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12:690049, 2021.

[21] Md Sharear Saon, Kevin Boehm, Grace Fu, Ian Hou, Jerry Yu, Brent M Znosko, and Jie Hou. Exploring the efficiency of deep graph neural networks for rna secondary structure prediction. *bioRxiv*, pages 2024–10, 2024.

[22] Pietro Bongini, Niccolò Pancino, Franco Scarselli, and Monica Bianchini. Biognn: how graph neural networks can solve biological problems. In *Artificial Intelligence and Machine Learning for Healthcare: Vol. 1: Image and Data Analytics*, pages 211–231. Springer, 2022.

[23] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[24] Zohar Rosenwasser, Erez Levanon, Michael Levitt, and Gal Oren. Detection of rna editing sites by gpt fine-tuning. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024.

[25] Pei Zhang, Yuanzhen Zhu, Qunfei Guo, Ji Li, Xiaoyu Zhan, Hao Yu, Nianxia Xie, Huishuang Tan, Nina Lundholm, Lydia Garcia-Cuetos, et al. On the origin and evolution of rna editing in metazoans. *Cell Reports*, 42(2), 2023.

[26] Boris Knyazev, Graham W Taylor, and Mohamed Amer. Understanding attention and generalization in graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[27] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[28] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[29] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[30] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[31] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[32] Manato Akiyama and Yasubumi Sakakibara. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4(1):lqac012, 2022.

[33] Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 6(5):548–557, 2024.

[34] 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*, 21(12):2287–2298, 2024.

[35] Rafael Josip Penić, Tin Vlašić, Roland G Huber, Yue Wan, and Mile Šikić. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *Nature Communications*, 16(1):5671, 2025.

[36] Yichong Zhao, Kenta Oono, Hiroki Takizawa, and Masaaki Kotera. Generrna: A generative pre-trained language model for de novo rna design. *PLoS One*, 19(10):e0310814, 2024.

[37] Yijia Xiao, Edward Sun, Yiqiao Jin, and Wei Wang. Rna-gpt: Multimodal generative system for rna sequence understanding. *arXiv preprint arXiv:2411.08900*, 2024.

[38] Han Huang, Ziqian Lin, Dongchen He, Liang Hong, and Yu Li. Ribodiffusion: tertiary structure-based rna inverse folding with generative diffusion models. *Bioinformatics*, 40(Supplement_1):i347–i356, 2024.

[39] Kengo Yamada and Michiaki Hamada. Bert-rbp: Predicting rna–protein interactions using bert embeddings. *Bioinformatics*, 38(5):1235–1242, 2022.

[40] Yekaterina Shulgina, Marena I Trinidad, Conner J Langeberg, Hunter Nisonoff, Seyone Chithrananda, Petr Skopintsev, Amos J Nissley, Jaymin Patel, Ron S Boger, Honglue Shi, et al. Rna language models predict mutations that improve rna function. *Nature Communications*, 15(1):10627, 2024.

[41] He Zhang, Hailong Liu, Yushan Xu, Yiming Liu, Jia Wang, Yan Qin, Haiyan Wang, Lili Ma, Zhiyuan Xun, Timothy K Lu, et al. Deep generative models generate mrna sequences with enhanced translation capacity and stability. *bioRxiv*, pages 2024–06, 2024.

[42] Yupeng Li, Fan Wang, Jiaqi Yang, Zirong Han, Linfeng Chen, Wenbing Jiang, Hao Zhou, Tong Li, Zehua Tang, Jianxiang Deng, et al. Deep generative optimization of mrna codon sequences for enhanced protein production and therapeutic efficacy. *bioRxiv*, pages 2024–09, 2024.

[43] Huseyin Avni Tac, Mustafa Koroglu, and Ugur Sezerman. Rddsvm: accurate prediction of a-to-i rna editing sites from sequence using support vector machines. *Functional & Integrative Genomics*, 21(5):633–643, 2021.

[44] Zhangyi Ouyang, Feng Liu, Chenghui Zhao, Chao Ren, Gaole An, Chuan Mei, Xiaochen Bo, and Wenjie Shu. Accurate identification of rna editing sites from primitive sequence with deep neural networks. *Scientific Reports*, 8(1):6005, 2018.

[45] Xin Liu, Tao Sun, Anna Shcherbina, Qin Li, Inga Jarmoskaite, Kalli Kappel, Gokul Ramaswami, Rhiju Das, Anshul Kundaje, and Jin Billy Li. Learning cis-regulatory principles of adar-based rna editing from crispr-mediated mutagenesis. *Nature communications*, 12(1):2165, 2021.

[46] Yue Jiang, Lina R Bagepalli, Bora S Banjanin, Yiannis A Savva, Yingxin Cao, Lan Guo, Adrian W Briggs, Brian Booth, and Ronald J Hause. Generative machine learning of adar substrates for precise and efficient rna editing. *bioRxiv*, pages 2024–09, 2024.

[47] Jiandong Wang, Scott Ness, Roger Brown, Hui Yu, Olufunmilola Oyebamiji, Limin Jiang, Quanhu Sheng, David C Samuels, Ying-Yong Zhao, Jijun Tang, et al. Editpredict: Prediction of rna editable sites with convolutional neural network. *Genomics*, 113(6):3864–3871, 2021.

[48] Zohar Rosenwasser, Erez Levanon, Michael Levitt, and Gal Oren. Leveraging gpt continual fine-tuning for improved rna editing site prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

[49] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6:1–14, 2011.

[50] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.