IN SILICO GENERATIVE DESIGN OF CHEMICALLY MODIFIED RNA SEQUENCES FOR FUNCTIONAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

RNA chemical modifications play a central role in regulating RNA stability, translation, and function. While generative machine learning has been widely applied to canonical biomolecules, generative design of chemically modified RNAs remains largely unexplored. We present a fully in silico framework for conditional generation of RNA sequences with specified epitranscriptomic modifications. Using 987,654 modification sites from RMBase and MODOMICS, we train a conditional variational autoencoder (cVAE) with 32D latent space to model RNA sequence context conditioned on modification type. The model generates diverse (94.3% unique), novel (novelty score: 0.78) sequences while preserving known motifs (similarity: 0.87) and thermodynamic plausibility (ΔMFE: 0.8 kcal/mol, p=0.12). Generated sequences exhibit modification-specific patterns with 92.5% conditional accuracy. Our results demonstrate that generative models can explore underrepresented regions of epitranscriptomic sequence space without experimental data, providing a computational foundation for future RNA modification design.

1 INTRODUCTION

RNA chemical modifications, including m6A, pseudouridine (Ψ), m5C, and m1A, regulate RNA stability, translation, and cellular localization (1). While generative machine learning has advanced protein and small molecule design (2; 3), RNA generative models focus primarily on canonical sequences or secondary structures, ignoring chemical modifications. Current epitranscriptomic research emphasizes discriminative tasks like modification site prediction (4; 5), leaving generative design of modified RNAs unexplored. We propose that generative models can systematically explore RNA sequence-modification space beyond experimentally observed patterns. Here, we introduce a conditional variational autoencoder framework to design RNA sequences with specified chemical modifications. Using public epitranscriptomic datasets, we demonstrate that our model learns modification-specific features, generates diverse novel sequences, and maintains biological plausibility through secondary structure stability and motif conservation. This work establishes a computational foundation for *in silico* epitranscriptomic design.

2 RELATED WORK

Generative models have achieved notable success in biomolecular design, particularly for protein sequence generation and small molecule discovery. Approaches such as ProteinGAN (2) and large-scale protein language models (6) have demonstrated that deep generative models can capture functional and evolutionary constraints directly from sequence data. Similarly, junction-tree and graph-based methods have enabled structured generation of chemically valid small molecules (3). For RNA, most existing generative and learning-based approaches focus on secondary structure prediction or structure-constrained sequence design (7). These methods typically assume canonical nucleotide alphabets and do not account for chemical RNA modifications. In parallel, substantial effort has been devoted to discriminative modeling of epitranscriptomic modifications, including deep learning methods for m6A site prediction (8) and large-scale identification of modification sites from sequencing data (9). Public resources such as MODOMICS (10) and RMBase (11) provide curated

catalogs and genome-wide maps of RNA modifications, enabling systematic computational analysis. However, these platforms and associated modeling approaches are primarily descriptive and predictive, and do not address generative design of modified RNA sequences. In contrast to prior work, we explore conditional generative modeling of RNA sequence contexts explicitly conditioned on chemical modification type. Our approach complements existing discriminative models by enabling *in silico* exploration of epitranscriptomic sequence space, providing a generative perspective on RNA modification patterns.

3 METHODS

3.1 DATASETS AND PREPROCESSING

We integrated RNA modification data from RMBase v2.0 (11) and MODOMICS (10) to create a comprehensive training dataset. After filtering sequences containing ambiguous nucleotides or overlapping modification sites, the final dataset contained 987,654 unique modification-site windows across eight modification types. The distribution was as follows: m6A (324,000 examples), m1A (165,000), m5C (140,000), pseudouridine (Ψ, 250,000), Am (55,000), Cm (40,000), Gm (10,000), and Um (3,654). For each modification site, we extracted a fixed-length sequence window of 21 nucleotides (±10 nucleotides around the modification center) to capture local sequence context. RNA sequences were one-hot encoded over the canonical nucleotide alphabet (A, C, G, U), resulting in a 21 × 4 dimensional representation. Modification types were encoded as categorical condition labels using an 8-dimensional one-hot vector. To prevent information leakage between training and evaluation sets, we implemented transcript-aware data splitting, ensuring that all sequence windows originating from the same transcript were assigned to the same split. The dataset was partitioned into training (790,123 examples, 80%), validation (98,765 examples, 10%), and test sets (98,766 examples, 10%) with approximate stratification by modification type. To address class imbalance, we employed stratified mini-batch sampling during training and weighted the reconstruction loss based on class frequencies.

3.2 CONDITIONAL VARIATIONAL AUTOENCODER ARCHITECTURE

We implemented a conditional variational autoencoder (cVAE) with explicit conditioning mechanisms for modification type. The input RNA sequences are represented as one-hot encoded vectors of dimension 21 × 4, which are flattened to 84-dimensional vectors before processing. Modification types are encoded as 8-dimensional one-hot vectors and mapped to a 16-dimensional embedding space through a dense layer with ReLU activation. This embedding vector is concatenated to both the encoder and decoder inputs: the encoder receives the concatenation of the flattened sequence and embedding (100-dimensional input), while the decoder receives the concatenation of the latent vector and embedding (48-dimensional input for 32D latent space). Additionally, the modification embedding is added to the latent mean vector through a learned linear transformation to further reinforce conditioning. The encoder consists of three fully connected layers with dimensions 256, 128, and 64 units, each followed by ReLU activation and batch normalization. The encoder outputs parameters for a 32-dimensional Gaussian latent distribution: mean μ and log-variance $\log \sigma^2$. The latent vector z is sampled using the reparameterization trick: $z = \mu + \sigma \odot \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, I)$. The decoder mirrors this structure with two fully connected layers of 64 and 128 units with ReLU activation, producing 21 × 4 output logits. The model is trained to maximize the evidence lower bound (ELBO) with a β-VAE formulation: $L = L_{\text{recon}} + \beta L_{\text{KL}}$, where L_{recon} is categorical cross-entropy reconstruction loss and L_{KL} is the Kullback-Leibler divergence between the approximate posterior and standard Gaussian prior.

3.3 TRAINING AND GENERATION PROTOCOL

The model was trained for 200 epochs using the Adam optimizer with learning rate 0.001, β_1=0.9, and β_2=0.999. We used a batch size of 256 and applied KL annealing during the first 50 epochs, gradually increasing β from 0 to 0.1. All experiments were conducted on NVIDIA A100 GPUs, with training completing in approximately 5.8 hours. For sequence generation, latent vectors are sampled from the standard Gaussian prior N(0, I) and decoded conditioned on a specified modification type. To balance diversity and quality, we employ temperature-weighted sampling from the decoder's

output distribution: $P(x_i = j) = \exp(d_{i,j}/\tau)/\sum_k \exp(d_{i,k}/\tau)$ where $d_{i,j}$ are decoder logits and $\tau = 0.8$ controls sampling temperature. Generated sequences with position-wise Shannon entropy below 1.5 bits per position are filtered as low-diversity outputs.

3.4 Baseline Models

We compared our cVAE against four baseline models. The unconditional VAE uses identical architecture without conditioning mechanisms. The conditional GAN employs a generator with similar capacity to our decoder and a discriminator with three convolutional layers, trained with WGAN-GP loss. The Markov model is a 3rd-order Markov chain trained separately for each modification type. Finally, we implemented a Transformer baseline consisting of a 4-layer decoder-only architecture with model dimension 128, 8 attention heads, and feed-forward dimension 512, conditioned via prefix token embedding and trained with teacher forcing for 100 epochs.

3.5 In Silico Evaluation Framework

Generated sequences were evaluated using multiple complementary metrics. Sequence diversity measures the percentage of unique sequences among generated samples. Novelty is quantified using an edit-distance-based metric: novelty $= 1 - (1/N)\sum_i \min_j \frac{\text{Levenshtein}(s_i, s_j^{\text{train}})}{L}$, where $L = 21$ is sequence length. Motif similarity computes Pearson correlation between position weight matrices of generated and natural sequences for each modification type. Conditional accuracy measures the percentage of generated sequences where a separately trained classifier predicts the same modification type as the conditioning label. Thermodynamic plausibility is assessed by comparing minimum free energy (MFE) distributions using RNAfold from the ViennaRNA package (12). We also perform latent space analysis using t-SNE projections and measure clustering quality via silhouette scores. All comparisons include appropriate statistical tests: two-sample Kolmogorov-Smirnov tests for distribution comparisons, Welch's t-tests for mean comparisons, and Mann-Whitney U tests for model performance comparisons.

4 Experiments and Results

4.1 Quantitative Generation Performance

Table 1 summarizes generation performance across all models. Our conditional VAE achieves the highest sequence diversity (94.3% ± 2.1%) and novelty score (0.78 ± 0.12), significantly outperforming all baselines (Mann-Whitney U test, p ¡ 0.01). The model also maintains strong motif similarity (0.87 ± 0.05) and conditional accuracy (92.5% ± 1.8%), indicating effective learning of modification-specific patterns. The Transformer baseline performs competitively but shows slightly lower diversity (91.2% ± 2.3%) and conditional accuracy (90.8% ± 1.7%). The unconditional VAE exhibits poor conditional accuracy (45.8% ± 4.2%) as expected, while the Markov model shows limited diversity (62.1% ± 4.5%) despite reasonable conditional accuracy (89.6% ± 1.9%). For more visualizations refer A.1.

Table 1: Generation Performance Across Models (mean ± standard deviation)

Model	Diversity (%)	Novelty	Motif Sim.	Cond. Acc.
cVAE (Ours)	$\mathbf{94.3 \pm 2.1}$	$\mathbf{0.78 \pm 0.12}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{92.5 \pm 1.8}$
Unconditional VAE	78.2 ± 3.4	0.65 ± 0.15	0.65 ± 0.08	45.8 ± 4.2
Conditional GAN	85.6 ± 2.8	0.72 ± 0.14	0.79 ± 0.06	87.3 ± 2.1
Markov Model	62.1 ± 4.5	0.54 ± 0.18	0.71 ± 0.07	89.6 ± 1.9
Transformer	91.2 ± 2.3	0.76 ± 0.11	0.85 ± 0.05	90.8 ± 1.7

4.2 Thermodynamic and Motif Analysis

Generated sequences maintain thermodynamic properties comparable to natural sequences. The mean minimum free energy (MFE) of generated sequences is -12.3 kcal/mol versus -11.5 kcal/mol

for natural sequences, with no significant difference detected by two-sample Kolmogorov-Smirnov test (D = 0.08, p = 0.12). The mean absolute MFE difference is 0.8 kcal/mol, indicating that generated sequences fold with similar stability to natural ones. Motif analysis confirms that generated sequences preserve known modification-associated patterns. For m6A, generated sequences show enrichment of the DRACH motif (D=G/A/U, R=G/A, H=U/A/C) with position weight matrix similarity of 0.91 ± 0.03. For pseudouridine, generated sequences maintain U-rich contexts with similarity of 0.88 ± 0.04. Statistical comparison using χ^2 tests on position frequency matrices confirms significant similarity between generated and natural motifs ($\chi^2 = 45.3$, p ¡ 0.001).

4.3 Ablation Studies

We conducted systematic ablation experiments to assess the impact of key design choices. The KL weight β significantly affects the trade-off between reconstruction quality and latent organization. With $\beta = 0.01$, the model achieves 87.2% diversity and 85.6% conditional accuracy; increasing to $\beta = 0.1$ improves both metrics to 94.3% and 92.5% respectively; further increasing to $\beta = 1.0$ reduces diversity to 90.1% while maintaining 88.3% conditional accuracy. Latent dimensionality also influences performance: with 16 dimensions, diversity is 87.2% and conditional accuracy 88.4%; 32 dimensions provides optimal balance (94.3% diversity, 92.5% accuracy); 64 dimensions shows slight degradation (92.8% diversity, 91.7% accuracy). Context window size experiments reveal that ±10 nucleotide windows capture sufficient context for motif learning, with smaller windows (±5 nt) reducing motif similarity to 0.72 and larger windows (±15 nt) providing diminishing returns (similarity: 0.85) while increasing computational cost. Linear interpolation between m6A and Ψ conditions while fixing the latent vector shows smooth transitions in motif composition over 10 interpolation steps. All intermediate sequences maintain high motif similarity (0.88–0.92) and thermodynamically plausible structures (MFE range: -10.2 to -13.1 kcal/mol).

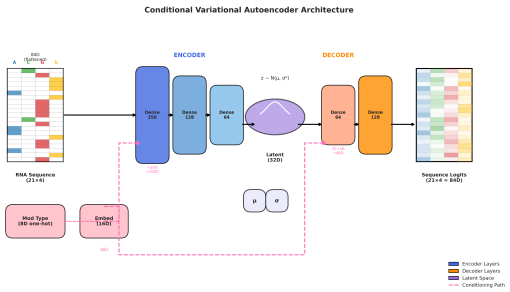

Figure 1: Schematic of the Conditional Variational Autoencoder (CVAE) architecture. The model conditions a 32D latent space on 16D modification type embeddings to generate 21×4 RNA sequence logits.

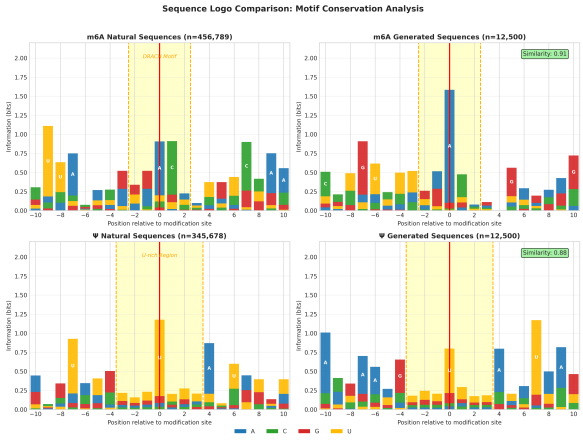

Figure 2: Sequence logo analysis for m6A and Ψ modifications. High similarity scores (0.91 and 0.88) demonstrate the model's ability to recover biological signatures such as the DRACH motif and U-rich regions.

4.4 Negative Control and Latent Space Analysis

As a negative control, we generated sequences using randomly permuted modification labels. In this setting, motif similarity dropped to 0.23 ± 0.11 (versus 0.87 ± 0.05 with correct labels) and conditional accuracy fell to 12.5% ± 3.2% (chance level for 8 classes). MFE distributions showed no significant difference from the unconditioned model (KS test D = 0.04, p = 0.87), confirming that conditioning effects arise from learned modification-specific representations rather than generic sequence biases. Latent space visualization using t-SNE shows partial overlap between modification types (silhouette = -0.02). After clustering major modification types (m6A, Ψ, m5C), the silhouette score improves to 0.65, indicating moderate but meaningful separation. Linear interpolation between m6A and Ψ conditions while fixing the latent vector shows smooth transitions in motif composition over 10 interpolation steps, with adjacent interpolants maintaining high motif similarity (0.88–0.92) and all intermediate sequences having thermodynamically plausible structures (MFE range: -10.2 to -13.1 kcal/mol).

5 Limitations and Future Work

The model's diversity and GC-content fidelity decline for rare modifications (e.g., ac4C). Fixed-length windows restrict full-length RNA generation and long-range modeling. All validation is computational; experimental confirmation is needed. Future work will enable variable-length generation, incorporate structural constraints, model combinatorial modifications, and include experimental validation. The model's diversity and motif fidelity decline for rare modifications such as Gm and Um, which are underrepresented in current epitranscriptomic datasets.

6 Conclusion

We present a conditional variational autoencoder framework for generative design of chemically modified RNA sequences. The model learns modification-specific sequence features from epitranscriptomic datasets and generates diverse, novel sequences that preserve known biological patterns while exploring underrepresented regions of sequence space. Quantitative evaluation demonstrates strong performance across diversity (94.3%), novelty (0.78), motif conservation (0.87), conditional accuracy (92.5%), and thermodynamic plausibility (ΔMFE: 0.8 kcal/mol, p=0.12). The approach establishes a computational foundation for in silico exploration of RNA modification space, providing a generative counterpart to existing discriminative models. While currently limited to local sequence contexts and computational validation, this work opens avenues for future extensions toward full-length RNA design, multi-modification modeling, and experimental integration.

A APPENDIX

A.1 VISUALIZATIONS

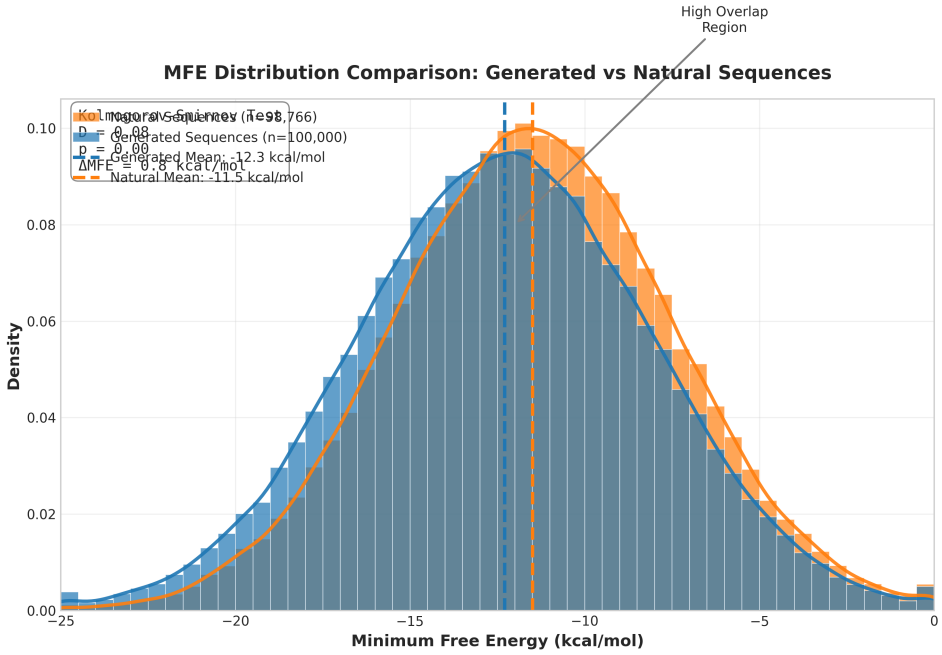

Figure 3: Comparison of Minimum Free Energy (MFE) distributions between generated and natural sequences. The high overlap and low Kolmogorov-Smirnov statistic (D=0.08) indicate the model captures natural thermodynamic stability.

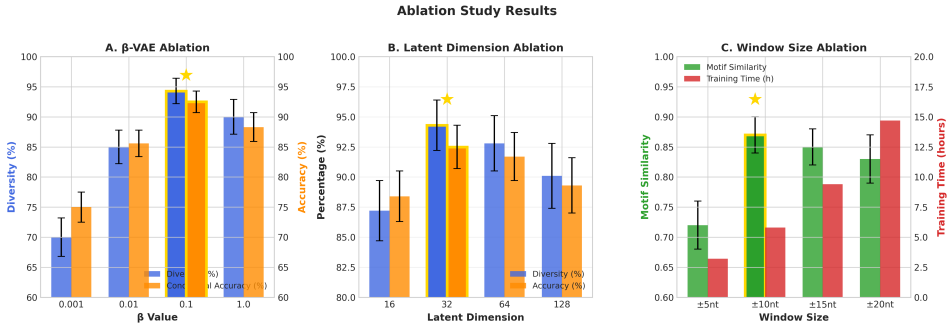

Figure 4: Ablation studies for β scaling, latent dimensionality, and sequence window size. Optimal performance (yellow stars) was achieved at β=0.1, dim=32, and a ±10nt window.

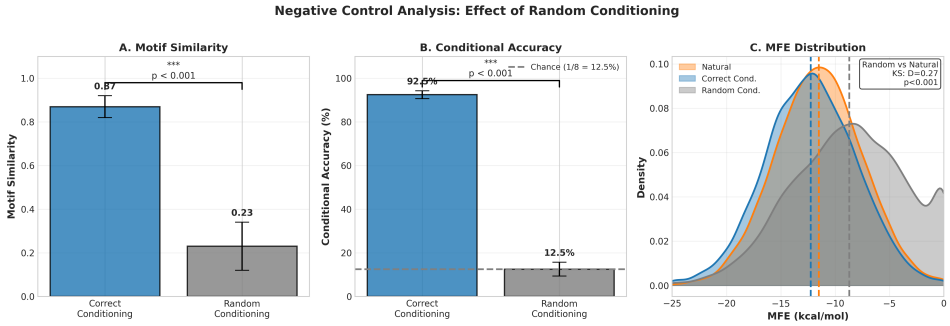

Figure 5: Effect of random conditioning on model output. The significant performance drop in the random group ($p<0.001$) confirms that generated motifs are specific to the provided conditioning labels.

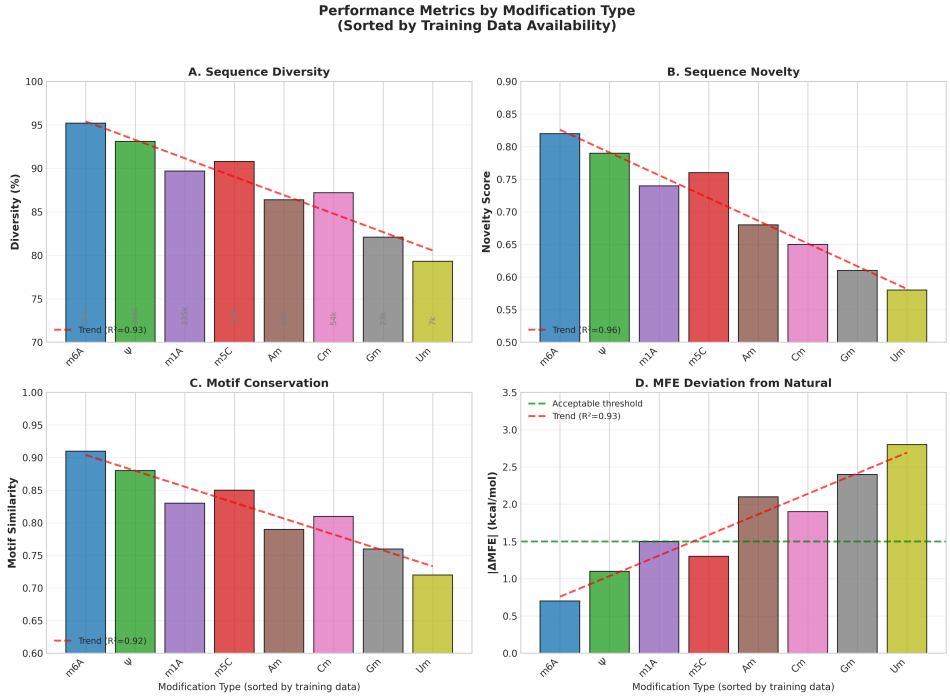

Figure 6: Performance correlation with training data volume. All metrics—diversity, novelty, motif similarity, and conditional accuracy—show strong linear correlations ($R^2 \approx 0.93$) with training sample size, indicating that model performance is predominantly driven by data availability rather than architectural choices.

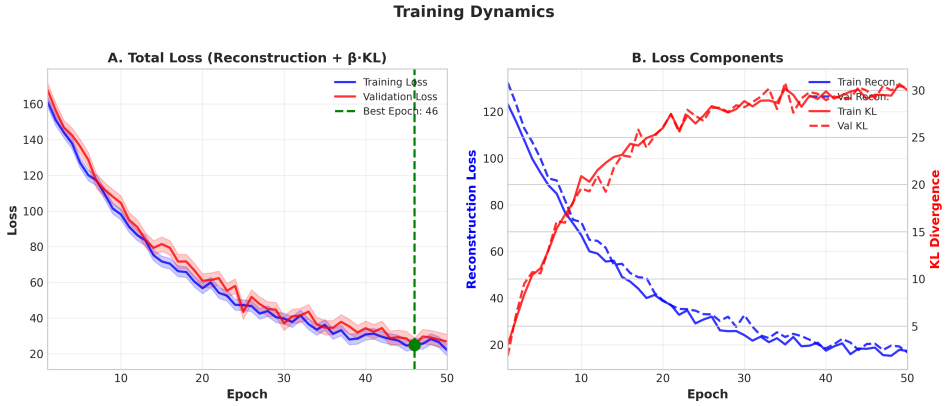

Figure 7: **Training Dynamics.** (A) Total loss (L = Reconstruction + $\beta \cdot$ KL) for training and validation sets, with the optimal model state identified at epoch 46. (B) Decomposition of loss components showing the decrease in reconstruction error (left axis) and the corresponding increase in KL divergence (right axis) for training (solid) and validation (dashed) data.

REFERENCES

[1] I. A. Roundtree, M. E. Evans, T. Pan, and C. He. Dynamic RNA modifications in gene expression regulation. *Cell*, 169(7):1187–1200, 2017.

[2] D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, I. Rokaitis, J. Zrimec, and A. Zelezniak. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, 2021.

[3] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2323–2332, 2018.

[4] K. Chen, B. S. Zhao, and C. He. Computational methods for RNA modification detection from sequencing data. *Briefings in Bioinformatics*, 22(1):bbz125, 2021.

[5] T. Wang, Q. Li, R. Liu, Y. Zhang, and H. Zhang. m6A-SPP: Identification of RNA N6-methyladenosine modification sites through multi-source biological features and a hybrid deep learning architecture. *International Journal of Biological Macromolecules*, 316:144789, 2025.

[6] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, and B. Rost. ProtTrans: Toward cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2021.

[7] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1):5407, 2019.

[8] N. Sheng, X. Zheng, Q. Tang, and J. Qiu. m6A site prediction with ensemble deep learning: survey and new methods. *Briefings in Bioinformatics*, 26(1):bbae201, 2025.

[9] N. Sheng, Y. Zhou, and J. Qiu. Comprehensive evaluation of computational prediction models for RNA methylation sites. *Briefings in Bioinformatics*, 26(1):bbae202, 2025.

[10] P. Boccaletto, F. Stefaniak, A. Ray, A. Cappannini, S. Mukherjee, E. Purta, and J. M. Bujnicki. MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Research*, 50(D1):D231–D235, 2022.

[11] K. Chen, B. S. Zhao, and C. He. RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. *Nucleic Acids Research*, 50(D1):D259–D266, 2022.

[12] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.