THERMOTOKEN: THERMODYNAMIC RESPONSE-CONDITIONED ADAPTATION FOR RNA SECONDARY STRUCTURE PREDICTION UNDER FAMILY SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-family generalization remains challenging for RNA secondary-structure prediction. RNA foundation models provide transferable sequence representations, while thermodynamic models characterize folding through sequence-dependent free energies. Existing neuro-physical approaches generally evaluate thermodynamic information at a fixed parameterization, leaving the local variation of pairing preferences around that state unobserved. We show that controlled perturbations of nearest-neighbor parameters expose how probability mass redistributes among competing folds, providing structure information beyond a single base-pair probability (BPP) map. Building on this observation, we introduce ThermoToken, which constructs a thermodynamic response representation from perturbed BPP maps and combines it with RiNALMo features and the current pairing state to iteratively revise secondary structure. Repeated updates progressively adjust pair assignments as the predicted structural configuration evolves. On strict family-held-out Archive-II evaluation, ThermoToken raises family-macro flexible F1 from 0.7193 to 0.7455, with improvements across all nine target families. More reliable prediction across RNA families can support RNA target characterization and structure-guided molecular design when annotated examples from the target family are limited.

1 INTRODUCTION

RNA secondary-structure prediction remains difficult under family shift, where sequence regularities learned from observed families often transfer poorly to unseen ones (Szikszai et al., 2022). Large pretrained RNA models provide transferable representations, yet pair prediction remains driven by sequence-derived evidence whose reliability varies across families (Chen et al., 2022; Penić et al., 2025; Zou et al., 2024). Thermodynamic models provide complementary energetic information, but existing neuro-physical approaches generally evaluate it at a single parameterization (Sato et al., 2021; Zhu et al., 2025). A single base-pair probability (BPP) map therefore describes one operating point of the folding ensemble. Controlled perturbations reveal how nearby energetic changes redistribute support among competing pairs.

For an RNA sequence x, a learned predictor assigns a score $s(Y \mid x)$ to a candidate structure Y, whereas a thermodynamic model defines a folding ensemble $p_\vartheta(Y \mid x) \propto \exp[-\Delta G_\vartheta(Y \mid x)/RT]$. A BPP map marginalizes this ensemble into pairwise support. At a reference parameterization ϑ_0, it records pair occupancy under one energetic condition. Nearby perturbations reweight competing folds, so pairs with similar reference probabilities can exhibit different local responses. The finite difference $P(x; \vartheta_0 + \delta) - P(x; \vartheta_0)$ measures this sensitivity around the reference state. The systematic analysis of Zuber et al. (2017) showed across 1,663 RNAs that perturbing the 291 independent Turner-2004 parameters can substantially alter both BPPs and predicted structures.

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

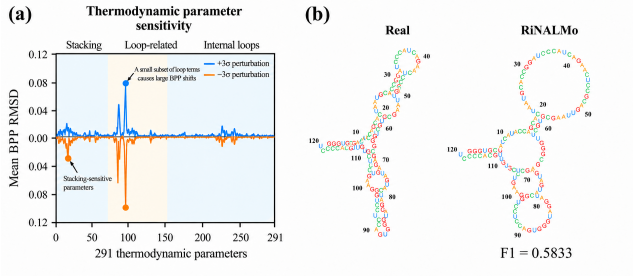


Figure 1: **Motivation. (a)** BPP sensitivity across 291 thermodynamic parameters, showing heterogeneous pairing responses. **(b)** Reference and RiNALMo-predicted 5S rRNA. Missing pairs enlarge the loop and disrupt local stem organization.

Figure 1a shows heterogeneous changes in pair support under nearby thermodynamic perturbations. Figure 1b shows a 5S rRNA in which missed stem-closing pairs enlarge the loop and alter local stem organization. These observations meet at partner competition. Perturbations expose how candidate support changes around the reference condition, while a missing pair changes the alternatives available to both endpoints. Once the pairing state changes, a later correction faces a different set of competing interactions.

In the nearest-neighbor model used by ViennaRNA (Lorenz et al., 2011), the free energy of a candidate structure is assembled from local energetic contributions. For a candidate structure Y,

$$\Delta G_\vartheta(Y \mid x) = \sum_{c \in \mathcal{C}(Y,x)} g_c(\vartheta). \tag{1}$$

The corresponding partition function is

$$Z_\vartheta(x) = \sum_Y \exp\left[-\frac{\Delta G_\vartheta(Y \mid x)}{RT}\right], \tag{2}$$

and the probability of a candidate pair is obtained by marginalizing over structures that contain it,

$$P_\vartheta(i,j \mid x) = \frac{1}{Z_\vartheta(x)} \sum_{Y:(i,j)\in Y} \exp\left[-\frac{\Delta G_\vartheta(Y \mid x)}{RT}\right]. \tag{3}$$

ViennaRNA evaluates these ensemble quantities through dynamic-programming recursions without enumerating the structure space. A perturbation changes relative free energies and Boltzmann weights and hence the marginal support assigned to individual pairs. The resulting response characterizes pair competition around the reference condition. The current pairing state supplies complementary context because changing one assignment changes the alternatives available to nearby nucleotides.

ThermoToken combines these two sources of information. BPP maps recomputed under controlled perturbations form a thermodynamic response representation, which is combined with RNA foundation-model features and the current pairing state in a shared correction operator. Each update revises the score field presented to the next pass, where pair competition is evaluated under the updated structural context. Under strict family separation on Archive-II, ThermoToken improves family-macro F1 from 0.7193 to 0.7455, with gains across all nine held-out RNA families. Performance reaches 0.9216 on 5S rRNA, while Telomerase RNA shows the largest family-wise improvement of 7.19 percentage points.

Contributions.

- We introduce thermodynamic response as a pairwise representation of how BPP support changes under controlled perturbations of RNA folding energetics.
- We develop ThermoToken, which combines thermodynamic response with the current pairing state and revises coupled pair assignments through successive updates.
- We demonstrate gains across all nine held-out Archive-II families and separate the contributions of thermodynamic response and iterative correction.

2 Related Work

2.1 RNA Secondary Structure Prediction under Family Shift

Task-specific predictors such as SPOT-RNA, E2Efold, UFold, RNAformer, and DEPfold learn base-pairing patterns directly from sequence (Singh et al., 2019; Chen et al., 2020; Fu et al., 2022; Franke et al., 2024; Wang & Cohen, 2025). RNA language models, including RNA-FM and RiNALMo, pretrain sequence representations on large unlabeled RNA corpora and transfer them to structure prediction (Chen et al., 2022; Penić et al., 2025). Family-held-out evaluation tests a stricter form of generalization by excluding structurally related target-family RNAs from training and validation (Szikszai et al., 2022). We adopt this setting with a frozen RiNALMo backbone and study thermodynamic response as sequence-specific evidence for cross-family transfer.

2.2 Thermodynamic Modeling for RNA Structure Prediction

Thermodynamic folding algorithms estimate RNA secondary structure from nearest-neighbor free-energy parameters. RNAstructure and ViennaRNA provide widely used implementations of thermodynamic folding and partition-function calculations (Reuter & Mathews, 2010; Lorenz et al., 2011). CONTRAfold provides a discriminatively trained probabilistic alternative without physics-based thermodynamic parameters (Do et al., 2006). Hybrid methods incorporate energetic information into learned predictors. MXfold2 integrates a nearest-neighbor thermodynamic model during training, and BPfold supplies a base-pair motif energy map as an auxiliary thermodynamic prior (Sato et al., 2021; Zhu et al., 2025). In both methods, the thermodynamic information enters the predictor in a fixed form for a given sequence. Parameter-sensitivity analysis addresses a related question by measuring how predicted structures or pairing probabilities vary when individual free-energy terms are perturbed (Zuber et al., 2017). ThermoToken brings this local response into RiNALMo score correction, extending learned prediction beyond a single reference thermodynamic state.

3 Method

3.1 Problem Formulation and Overall Framework

Given an RNA sequence $x = (x_1, \ldots, x_L)$, we predict a pairing matrix $Y \in \{0,1\}^{L\times L}$, where $Y_{ij} = 1$ denotes a base pair between nucleotides i and j. We study family-held-out prediction with the target family excluded from training and validation. RiNALMo supplies nucleotide representations and initial pair scores. Thermodynamic response provides pairwise features used to revise these scores.

Figure 2 summarizes the model. The frozen RiNALMo encoder and structure head produce nucleotide features and initial pair logits. In parallel, the thermodynamic branch computes BPP maps under the reference parameterization and controlled perturbations. A shared correction operator receives the current pair scores, sequence features, and thermodynamic representation and predicts a residual update.

For sequence x, the frozen encoder produces $H(x) = \mathrm{Encoder}_{\mathrm{RiNALMo}}(x) \in \mathbb{R}^{L\times d}$ and the frozen structure head produces initial pairing logits $S^0(x) = \mathrm{Head}_{\mathrm{RiNALMo}}(H(x)) \in \mathbb{R}^{L\times L}$. The thermodynamic branch computes $\Phi(x)$ from the same sequence without structure labels. At update t,

$$\Delta S^t(x) = F_\psi\big(S^t(x), H(x), \Phi(x)\big), \qquad S^{t+1}(x) = S^t(x) + \Delta S^t(x). \tag{4}$$

After three updates, $S^3(x)$ is passed to the sequence-aware constrained decoder used by the RiNALMo baseline. $H(x)$ and $\Phi(x)$ are computed once. The pairing state evolves from S^0 to S^3. The residual form preserves the backbone scores as the starting point and concentrates learned adaptation on pair decisions that require revision.

3.2 Thermodynamic Response Representation

Let ϑ_0 denote the reference Turner parameterization (Mathews et al., 2004) and $\mathcal{G} = \{\mathrm{stack}, \mathrm{hairpin}, \mathrm{mismatch}, \mathrm{terminal}\}$. We organize related nearest-neighbor terms into these four perturbation groups. The compact grouping is motivated by the heterogeneous parameter effects

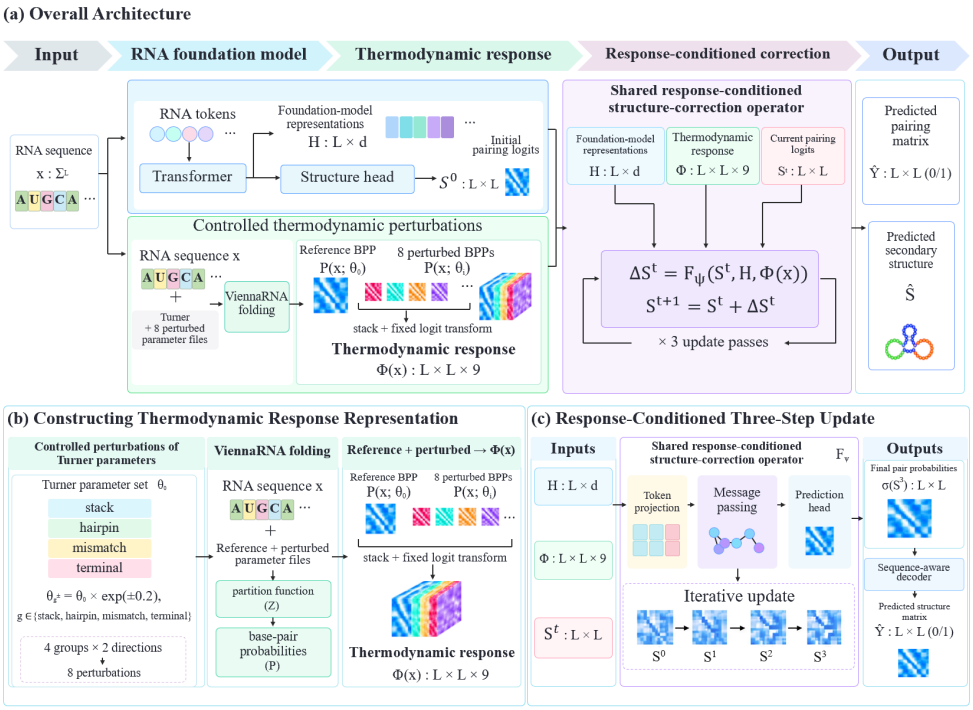


Figure 2: **ThermoToken architecture.** **(a)** Overall prediction pipeline. **(b)** Thermodynamic response construction from reference and perturbed BPP maps. **(c)** Three iterative residual updates of the pairing state.

identified by Zuber et al. (2017). For each $g \in \mathcal{G}$, the associated energy parameters are scaled by $\exp(+0.2)$ and $\exp(-0.2)$ with the other groups held at their reference values, giving ϑ_g^+ and ϑ_g^-. ViennaRNA recomputes the partition function and BPPs for every parameterization (Lorenz et al., 2011). With $P(x;\vartheta) \in [0,1]^{L\times L}$ denoting the BPP matrix,

$$\widetilde{\Phi}(x) = \Big[P(x;\vartheta_0), \{P(x;\vartheta_g^+), P(x;\vartheta_g^-)\}_{g\in\mathcal{G}}\Big] \in [0,1]^{L\times L\times 9}. \tag{5}$$

Each channel is converted to stabilized log-odds, yielding $\Phi(x) \in \mathbb{R}^{L\times L\times 9}$. The perturbations are fixed before training and use no target-family labels.

The common perturbation magnitude places the four energy groups on a shared local scale. Each BPP map reflects both the scaled energy terms and their effect on the folding ensemble. Grouping related Turner terms keeps the representation compact while distinguishing stacking, hairpin, mismatch, and terminal contributions.

The nine channels place reference occupancy beside positive and negative responses for each candidate pair. Both directions are retained because responses can be asymmetric around the reference state. Pairs with similar reference probabilities may differ in their sensitivity or exchange support with alternative partners. In the $L \times L$ pair space, the correction operator can compare these profiles directly with learned pair scores. Together, occupancy, response direction, and perturbation identity describe how each candidate pair sits within its nearby folding landscape.

3.3 Response-Conditioned Structure Correction

The correction operator F_ψ combines nucleotide features $H(x)$, current pair logits $S^t(x)$, and thermodynamic response $\Phi(x)$. The three sources remain distinct until pairwise message passing, allowing each score change to depend on sequence context, the current structural hypothesis, and thermodynamic behavior for the same candidate interaction.

Response-conditioned message passing. Nucleotide embeddings are projected to compact node states and updated with two message-passing layers. Pairwise messages use the endpoint states, current pair logits, and nine thermodynamic channels. An explicit unpaired state gives every nucleotide a valid alternative when all paired candidates receive weak support.

Partner competition arises because a nucleotide can have only one partner in the decoded structure. Each candidate pair contributes a message, conditioned on its current score and thermodynamic channels, to both endpoint states. Aggregating these messages summarizes the alternatives available at each nucleotide. The second layer propagates this revised context to neighboring interactions without changing the frozen backbone representation. A symmetric pairwise readout then gives the same correction under exchange of the two nucleotides.

Residual structure correction. A pairwise head combines the updated endpoint states with the current pair logit, thermodynamic features, sequence distance, and pair-competition features. Its symmetric output $\Delta S^t(x)$ adjusts individual pair scores.

Iterative update. We apply the shared operator for three updates, producing S^1, S^2, and S^3 from the backbone state S^0. Updating S^t changes the competition presented to the next update even though $H(x)$ and $\Phi(x)$ remain fixed. The same correction rule therefore acts on progressively revised score fields, allowing later updates to reconsider alternatives affected by earlier pair decisions. The update-depth choice is examined through the trajectory and saturation analysis in Section 4.5. Appendix B.5 reports sensitivity sweeps over T and λ_{BCE}.

3.4 Training and Inference

Algorithm 1 summarizes the procedure. Training supervises every updated state, and inference decodes the final logits S^3.

Algorithm 1 Response-conditioned three-step update

Require: RNA sequence x
$H(x) \leftarrow \mathrm{Encoder}_{\mathrm{RiNALMo}}(x)$
$S^0(x) \leftarrow \mathrm{Head}_{\mathrm{RiNALMo}}(H(x))$
$\Phi(x) \leftarrow \mathrm{ThermodynamicResponse}(x)$
$S \leftarrow S^0(x)$
for $t = 1, 2, 3$ **do**
  $Z \leftarrow \mathrm{ResponseMessagePassing}(H(x), S, \Phi(x))$
  $\Delta S \leftarrow \mathrm{PairwiseCorrection}(Z, S, \Phi(x))$
  $S \leftarrow S + \Delta S$
  if $t < 3$ and training **then**
    $S \leftarrow \mathrm{stopgrad}(S)$
  end if
end for
$\hat{Y} \leftarrow \mathrm{Decode}(S)$
return $\hat{Y}$

Let $\overline{\mathrm{BCE}}_{+}^{(t)}$ and $\overline{\mathrm{BCE}}_{-}^{(t)}$ denote the mean binary cross-entropy over valid positive and negative candidate pairs at update t. The per-update objective combines class-balanced BCE with a tolerant soft-F1 term,

$$\mathcal{L}_{\text{bal-BCE}}^{(t)} = \tfrac{1}{2}\left(\overline{\mathrm{BCE}}_{+}^{(t)} + \overline{\mathrm{BCE}}_{-}^{(t)}\right), \qquad \mathcal{L}_{\text{tol-F1}}^{(t)} = 1 - \mathrm{SoftF1}_{\pm 1}\big(S^t(x), Y\big). \tag{6}$$

The class-balanced BCE gives positive and negative candidate sets equal contribution at each update. The soft-F1 term uses the one-nucleotide endpoint tolerance of the primary evaluation metric, aligning the auxiliary objective with the flexible-pair evaluation. Averaging over the T supervised updates gives

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\left[\lambda_{\mathrm{BCE}}\mathcal{L}_{\text{bal-BCE}}^{(t)} + \mathcal{L}_{\text{tol-F1}}^{(t)}\right]. \tag{7}$$

Intermediate logits are detached between update steps. Each update is optimized from the state produced by its predecessor, with gradients stopped across the full update trajectory. Supervision

at every updated state makes each application of the shared operator directly predictive instead of using early updates only as latent computation. The training design matches the shared-operator formulation, with F_ψ learning to improve the pairing configuration presented at each stage. The RiNALMo encoder and original structure head remain frozen, so training is confined to the added correction operator.

The adopted configuration uses $T = 3$ and $\lambda_{\mathrm{BCE}} = 0.5$ throughout the family-held-out experiments. Group I intron, tRNA, and tmRNA development tasks are used for sensitivity analysis without access to any held-out target family. Appendix B.5 reports the complete update-depth and loss-weight sweeps.

4 Experiments

We evaluate cross-family performance, structural corrections, component contributions, and update dynamics. Our analysis moves from family-level accuracy to the structural consequences of pair revision, linking prediction gains to changes in stem and loop organization. Ablations and update trajectories then isolate the roles of thermodynamic response and iterative state evolution across successive corrections.

4.1 Experimental Setup

We use nine family-held-out Archive-II splits with the target family excluded from source training and validation (Sloma & Mathews, 2016; Szikszai et al., 2022). RiNALMo-Giga is the frozen backbone and matched baseline, and training updates only the ThermoToken correction operator (Penić et al., 2025). Flexible F1 with a one-nucleotide pairing tolerance is the primary metric. Archive-II Macro is the unweighted mean over the nine target families.

Each held-out family defines a separate transfer task, with model selection restricted to validation data from the remaining source families and the target-family test set reserved for final evaluation. RiNALMo and ThermoToken share the encoder, structure head, sequence-aware decoder, and validation-based threshold selection, differing only by the added correction pathway. Archive-II Macro weights the nine target families equally.

4.2 Overall Performance

Table 1 reports flexible F1 for each held-out family. Published RNAstructure, CONTRAfold, MXfold2, UFold, and RNA-FM values are from Table 2 of the bioRxiv v1 release of Zou et al. (2024). ThermoToken raises the Archive-II macro average from 0.7193 to **0.7455**, a gain of 2.62 percentage points over the matched RiNALMo backbone, with improvements on all nine target families.

Table 1: **Family-held-out performance on Archive-II.** Flexible F1 by held-out family. Macro is the unweighted mean across tasks. All entries are displayed to four decimal places; published three-decimal values are zero-padded for consistent presentation. Best results in each column are bold.

Method	5S	16S	23S	Grp I	SRP	tRNA	tmRNA	Telo.	RNase P	Macro
Thermodynamic / statistical										
RNAstructure	0.6300	0.5700	0.7300	0.5400	0.6300	0.7000	0.4300	0.5000	0.5500	0.5900
CONTRAfold	0.6300	0.5800	0.7100	0.6000	0.5500	0.7700	0.4900	0.5400	0.6300	0.6100
Deep / hybrid predictors										
MXfold2	0.5400	0.5500	0.6400	0.4500	0.5000	0.6400	0.4600	0.3400	0.5100	0.5100
UFold	0.5300	0.4100	0.4500	0.4500	0.2600	0.2600	0.4000	**0.8000**	0.4100	0.4400
RNA foundation models										
RNA-FM	0.5700	0.1400	0.1900	0.1600	0.2500	0.7900	0.2800	0.0700	0.3100	0.3100
RiNALMo	0.8843	0.7370	0.8454	0.6565	0.6997	0.9312	0.8007	0.1203	0.7984	0.7193
ThermoToken	**0.9216**	**0.7585**	**0.8513**	**0.6780**	**0.7249**	**0.9501**	**0.8145**	0.1922	**0.8184**	**0.7455**

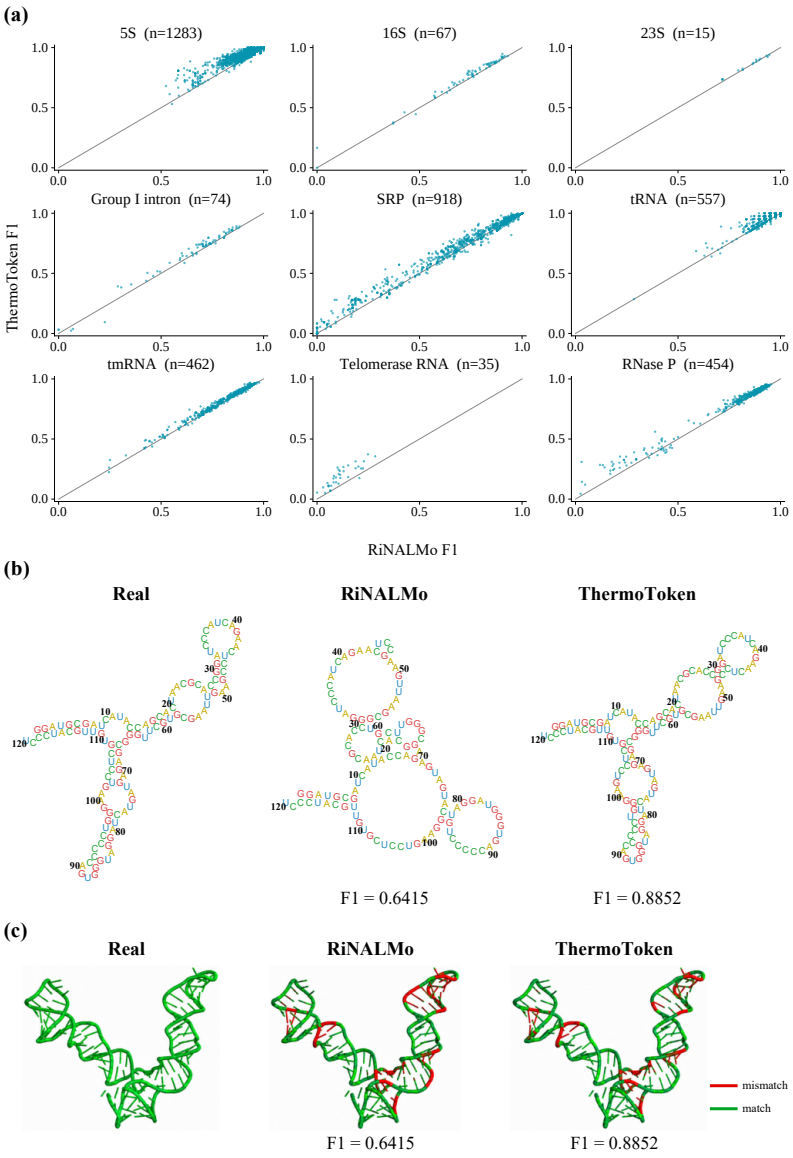


Figure 4: **Sequence-level and structural analysis. (a)** Per-RNA flexible F1 for ThermoToken and RiNALMo on the nine held-out families. **(b)** Secondary-structure prediction for a 120-nt 5S rRNA. **(c)** Three-dimensional view of the same base-pair differences.

The gains span markedly different backbone regimes. 5S rRNA rises from 0.8843 to 0.9216, while Telomerase RNA shows the largest family-wise improvement, from 0.1203 to 0.1922. The exact-pair macro F1 is 0.7305 in Appendix A, close to the 0.7455 flexible score, indicating that the improvement largely reflects near-exact pair recovery. We next examine how these changes appear in predicted structures.

4.3 Structural Analysis

Figure 4 summarizes sequence-level and structural effects. Panel (a) compares per-RNA flexible F1 across the nine held-out families, while panels (b) and (c) examine the same 120-nt 5S rRNA in two and three dimensions.

Figure 4a shows improvements across families and over a broad range of RiNALMo scores, matching the family-level trend. In Figure 4b, RiNALMo recovers much of the reference scaffold but leaves several interactions missing or misplaced. ThermoToken preserves that scaffold and recovers additional reference pairs, increasing flexible F1 from 0.6415 to **0.8852**. Figure 4c shows the same differences in three dimensions.

Stem closure links local pair assignments to adjacent loop boundaries, so coordinated pair recovery can change larger structural organization. Among 947 RNAs whose largest reference non-external loop contains at least 50 nucleotides, the mean base-pair deficit decreases by **18.30 pairs**, exact loop-location recovery rises from 5.1% to **40.7%**, and median loop IoU increases from 0.707 to 0.934. The broader 20-nt subset follows the same trend, with full statistics reported in Appendix A. These cohort measurements connect the pair-level gains to shifts in stem boundaries and loop placement.

4.4 Ablation Study

Table 2 separates thermodynamic response from iterative correction on the same three development families. All four settings use the same backbone, training protocol, and decoder.

Table 2: **Ablation of thermodynamic response and iterative updating.** Flexible F1 on three development families.

Input	Update	Group I	tRNA	tmRNA	Mean
None	Static	0.6339	0.9011	0.7367	0.7572
None	Iterative	0.6460	0.9188	0.7585	0.7744
Response	Static	0.6884	0.9193	0.7677	0.7918
Response	**Iterative**	**0.7109**	**0.9365**	**0.7783**	**0.8086**

Thermodynamic response adds **3.46 percentage points** with a static update and **3.42** with iterative updating, while iteration adds **1.72 percentage points** without response and **1.68** with it. The crossed effects vary by only 0.04 percentage points, and the summed isolated gains give 0.8090, close to 0.8086 for the full configuration. Response changes pair evidence and iteration changes the pairing configuration in which it is applied, yielding near-additive complementary contributions.

4.5 Response Across Evolving Pairing States

With thermodynamic response held fixed, Response Iterative adds 1.68 percentage points over Response Static in Table 2. Figure 3 traces the corresponding state evolution.

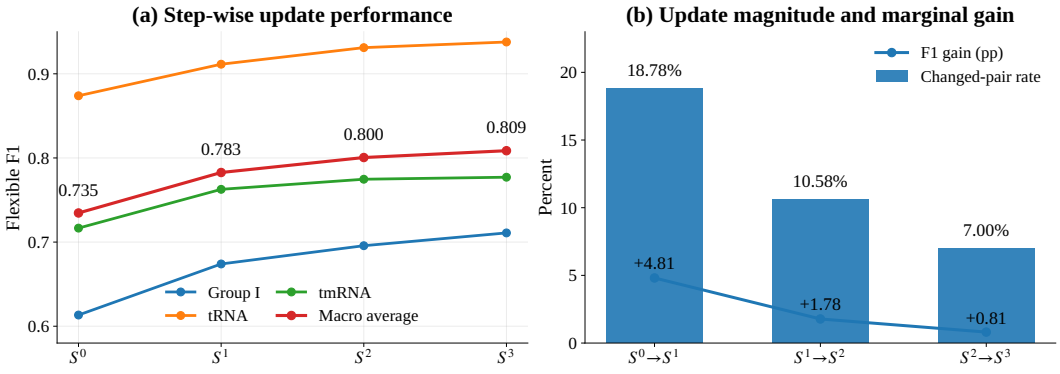


Figure 3: **Three-step update dynamics. (a)** Flexible F1 for S^0–S^3. **(b)** Fraction of changed pairs and incremental F1 gain at each update.

Figure 3 shows the largest pair revision and F1 gain in the first update. Later updates change fewer pairs and add smaller gains, while the $T = 4$ sweep in Appendix B.5 gives no further improvement.

Because the response representation is fixed, the trajectory reflects applying the same correction rule to successive pairing states. All three development families improve from S^0 to S^3, and the exact-pair and loop-level analyses follow the same direction. Performance stabilizes by the third update. The backbone features and thermodynamic response are computed once, so increasing T adds another evaluation of the correction operator rather than recomputing either branch. A fourth update therefore increases inference work after the pairing state has already reached its observed performance plateau. We use three updates at this convergence point.

4.6 Cross-Family Role of Thermodynamic Response

The experiments establish a sequence of evidence from local thermodynamic response to cross-family structural correction. Table 2 shows that response and iteration contribute separately. Thermodynamic response changes the evidence attached to each candidate pair, and repeated application evaluates that evidence under an evolving pairing configuration. Their near-additive gains place these effects at distinct stages of the correction process.

Figure 3 clarifies how this interaction develops. Most pair revisions occur in the first update, followed by progressively smaller changes as competing partners are resolved. The declining changed-pair rate and marginal F1 gain from S^1 to S^3, together with the $T = 4$ plateau, mark convergence of the shared operator. Three updates therefore retain the observed improvement with one fewer correction pass than four-step inference.

The structural analysis connects these local revisions to larger organization. Figure 4 shows pair recovery across different backbone score regimes and in the 5S rRNA example, and the loop cohort records corresponding changes in stem closure and loop placement. The perturbation axes keep a fixed thermodynamic meaning across sequences, whereas the induced BPP shifts reflect the folding ensemble of each RNA. This combination gives unseen-family sequences a shared physical coordinate system with sequence-specific measurements, linking perturbation-derived evidence to pair reassignment and structural reorganization.

5 Conclusion

ThermoToken introduces thermodynamic response as an additional coordinate system for RNA structure prediction under family shift. Controlled perturbations turn a fixed free-energy model into sequence-specific measurements of how pair support moves around its reference state, and a shared correction operator uses those measurements as the pairing configuration evolves. Across nine Archive-II held-out families, family-macro flexible F1 rises from 0.7193 to 0.7455, with gains in every target family. Structural and factorial analyses connect the improvement to pair recovery, stem closure, loop placement, and complementary contributions from response and iterative correction. The resulting framework shows how local sensitivity of a physical model can provide transferable evidence for RNA target characterization and structure-guided molecular design.

AI use statement

Generative AI tools were used during manuscript preparation for literature retrieval and organization, drafting and editing prose, formatting LaTeX and BibTeX, and checking consistency across claims, numbers, and citations. They were not used to generate or alter the experimental measurements reported in this paper. All AI-assisted manuscript content was reviewed and verified by the authors, who take responsibility for the final manuscript and reported results.

Reproducibility statement

The main text specifies the family-held-out evaluation protocol, matched backbone, and primary metric. Appendix B documents the deduplicated split construction, checkpoint and threshold selection, model configuration, training objective, optimization settings, ViennaRNA configuration, thermodynamic perturbation groups, decoding procedure, and sensitivity analysis. Appendix A reports supplementary structural and exact-pair evaluations. All reported experiments use the configurations described in the paper and appendix.

REFERENCES

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *BioRxiv*, pp. 2022–08, 2022.

Xinshi Chen, Yu Li, Ramzan Umarov, Xin Gao, and Le Song. Rna secondary structure prediction by learning unrolled algorithms. In *International Conference on Learning Representations*, 2020.

Chuong B Do, Daniel A Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

Jörg K. H. Franke, Frederic Runge, Ryan Köksal, Dominika Matus, Rolf Backofen, and Frank Hutter. Rnaformer: A simple yet effective model for homology-aware rna secondary structure prediction. *bioRxiv*, 2024. doi: 10.1101/2024.02.12.579881.

Laiyi Fu, Yingxin Cao, Jie Wu, Qinke Peng, Qing Nie, and Xiaohui Xie. Ufold: fast and accurate rna secondary structure prediction with deep learning. *Nucleic acids research*, 50(3):e14–e14, 2022.

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):26, 2011.

David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, 2004.

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.

Jessica S Reuter and David H Mathews. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129, 2010.

Kengo Sato, Manato Akiyama, and Yasubumi Sakakibara. Rna secondary structure prediction using deep learning with thermodynamic integration. *Nature communications*, 12(1):941, 2021.

Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications*, 10(1):5407, 2019.

Michael F Sloma and David H Mathews. Exact calculation of loop formation probability identifies folding motifs in rna secondary structures. *RNA*, 22(12):1808–1818, 2016.

Marcell Szikszai, Michael Wise, Amitava Datta, Max Ward, and David H Mathews. Deep learning models for rna secondary structure prediction (probably) do not generalize across families. *Bioinformatics*, 38(16):3892–3899, 2022.

Ke Wang and Shay B Cohen. Depfold: Rna secondary structure prediction as dependency parsing. In *International Conference on Learning Representations*, volume 2025, pp. 9800–9821, 2025.

Heqin Zhu, Fenghe Tang, Quan Quan, Ke Chen, Peng Xiong, and S Kevin Zhou. Deep generalizable prediction of rna secondary structure via base pair motif energy. *Nature Communications*, 16(1): 5856, 2025.

Shuxian Zou, Tianhua Tao, Sazan Mahbub, Caleb N Ellington, Robin Algayres, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P Xing. A large-scale foundation model for rna function and structure prediction. *bioRxiv*, pp. 2024–11, 2024.

Jeffrey Zuber, Hongying Sun, Xiaoju Zhang, Iain McFadyen, and David H Mathews. A sensitivity analysis of rna folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on rna secondary structure prediction. *Nucleic acids research*, 45(10): 6168–6176, 2017.

A Additional Evaluation

A.1 Large-Loop Structural Analysis

For each RNA, we define the largest loop as the largest non-external loop and compare predicted pairing density, loop size, and loop location with the reference structure. Among 3,860 RNAs with a definable largest non-external loop, Table S1 reports reference-loop subsets of at least 20 nt ($n = 1{,}037$) and 50 nt ($n = 947$). Pair-count difference and paired-nucleotide fraction bias are defined as prediction minus reference, so values closer to zero indicate smaller systematic deviation.

Table S1: **Large-loop structural fidelity.** RiNALMo and ThermoToken on two reference-loop size subsets. MAE denotes mean absolute error. IoU denotes intersection over union. pp denotes percentage points.

	$\geq$ 20 nt		$\geq$ 50 nt	
Metric	RiNALMo	ThermoToken	RiNALMo	ThermoToken
Pair-count difference	-25.46	**-7.42**	-24.91	**-6.61**
Paired-nt fraction bias (pp)	-13.86	**-3.93**	-13.60	**-3.56**
Largest-loop size MAE (nt)	102.44	**77.43**	109.73	**75.25**
Exact location match (%)	4.7	**37.3**	5.1	**40.7**
Median loop IoU	0.657	**0.896**	0.707	**0.934**
Median center distance (nt)	11.5	**4.0**	9.5	**1.5**

For tied largest loops, location recovery is counted as exact when any predicted largest-loop interval exactly matches a reference largest-loop interval.

A.2 Exact-Pair Evaluation

Table S2 reports exact-pair F1 alongside the primary flexible F1. The small difference between the two metrics for most families indicates that the reported performance is retained under exact pair matching.

Table S2: **ThermoToken flexible and exact-pair F1 by task.**

Test task	Flexible F1	Exact F1
5S rRNA	0.9216	0.9138
16S rRNA	0.7585	0.7519
23S rRNA	0.8513	0.8431
Group I intron	0.6780	0.6607
SRP RNA	0.7249	0.6899
tRNA	0.9501	0.9491
tmRNA	0.8145	0.7805
Telomerase RNA	0.1922	0.1834
RNase P	0.8184	0.8021

B Reproducibility Details

B.1 Task-Level Evaluation Configuration

Table S3 reports the deduplicated family-held-out split sizes and task-specific decision thresholds used for final decoding. Each row corresponds to a separately trained held-out-family task. ThermoToken is trained for five epochs. Model selection uses the nonzero epoch with the highest macro F1 over the source-family validation sets, with ties resolved in favor of the earlier epoch.

The source pool is deduplicated at the exact-sequence level before constructing the training and validation partitions, yielding sequence-disjoint source splits for every held-out task. Before deduplication, the corresponding training and validation partitions contained 45–75 exact sequence overlaps

depending on the held-out family. Checkpoint and decoding-threshold selection were repeated after deduplication. Deduplication did not change the selected checkpoint or decoding threshold for any held-out task. Exact sequence overlap between the source data and the held-out target-family test set is zero.

Table S3: **Deduplicated split sizes and decoding thresholds.**

Test task	Train	Validation	Test	Threshold
5S rRNA	2324	258	1283	0.07
16S rRNA	3419	379	67	0.04
23S rRNA	3465	385	15	0.23
Group I intron	3412	379	74	0.14
SRP RNA	2653	294	918	0.11
tRNA	2978	330	557	0.11
tmRNA	3063	340	462	0.07
Telomerase RNA	3447	383	35	0.22
RNase P	3070	341	454	0.09

B.2 Implementation Details

RiNALMo backbone. Each Archive-II held-out task uses the corresponding RiNALMo-Giga checkpoint fine-tuned for secondary-structure prediction. The checkpoint filename follows `rinalmo_giga_ss_archiveII-<task>_ft.pt`. The checkpoints are obtained from the official RiNALMo Zenodo release.[1] The experiments therefore use the task-specific secondary-structure checkpoints rather than initializing directly from `rinalmo_giga_pretrained.pt`. For both the matched RiNALMo baseline and ThermoToken, the RiNALMo encoder and secondary-structure prediction head are placed in evaluation mode and kept frozen throughout ThermoToken training.

Structure-correction operator. RiNALMo provides a 1280-dimensional nucleotide representation at each sequence position. ThermoToken projects these representations to 64-dimensional node states and combines them with the current $L \times L$ pair-score matrix and nine thermodynamic BPP channels. The correction operator contains two message-passing layers. Each message MLP has dimensions $192 \to 128 \to 64$. The symmetric pair head has dimensions $141 \to 128 \to 128 \to 1$. GELU activations are used in the trainable correction module, with no dropout. The correction operator contains 235,885 trainable parameters. Its parameters are shared across all update steps.

Optimization. Training processes one complete RNA at a time. We use AdamW with learning rate 3×10^{-4} and weight decay 0.01. ThermoToken is trained for five epochs with gradient-norm clipping at 1.0. No learning-rate scheduler is used. The optimizer is reinitialized at the beginning of epoch 4 with the same optimization hyperparameters. The main experiments use random seed 42. Model selection is based on macro F1 over the source-family validation sets, with ties resolved by selecting the earlier epoch.

ViennaRNA configuration. Thermodynamic quantities are computed with ViennaRNA 2.6.4 at 37°C using the built-in Turner-2004 nearest-neighbor parameterization. We use `dangles=2`, `noLP=False`, and `uniq_ML=True`. Partition-function calculations produce the reference and perturbed BPP matrices used by ThermoToken.

Thermodynamic perturbations. The response representation uses four selected groups of Turner-2004 parameters,

$$\mathcal{G} = \{\text{stack}, \text{hairpin}, \text{mismatch}, \text{terminal}\}.$$

For each group, the selected parameters are scaled jointly by $\exp(+0.2)$ and $\exp(-0.2)$ while all other thermodynamic parameters remain at their reference values. The four groups cover selected parameter classes rather than the complete Turner-2004 parameter set. Together with the unperturbed reference BPP map, the eight perturbed BPP maps form the nine thermodynamic channels used by ThermoToken.

[1] https://zenodo.org/records/15043668

Table S4 lists the parameter entries assigned to each perturbation group.

Table S4: **Thermodynamic perturbation groups.** Each group is perturbed independently in the positive and negative directions while all remaining Turner-2004 parameters remain fixed.

Group	ViennaRNA parameter entries
Stack	`stack` and `stack_enthalpies`
Hairpin	`hairpin` and `hairpin_enthalpies`
Mismatch	All `mismatch_*` parameter sections
Terminal	The two `TerminalAU` numerical entries in the `Misc` section

Final decoding. After the third update, ThermoToken applies a sigmoid to the corrected pair logits and uses the task-specific decision threshold associated with the corresponding frozen RiNALMo checkpoint. Pair scores are then passed to the same sequence-aware greedy decoder used by RiNALMo. Final decoding permits AU, UA, GC, CG, GU, and UG pairs, requires paired positions to be separated by at least four nucleotides, and assigns at most one partner to each nucleotide. The internal candidate mask used by the correction operator contains non-diagonal pairs between valid nucleotide positions. Canonical and wobble-pair constraints are imposed during final decoding.

B.3 Flexible-Pair Metric

Let Y and $\widehat{Y}$ be undirected reference and predicted base-pair sets. A predicted pair (i, j) matches a reference pair when either i agrees and j differs by at most one nucleotide, or j agrees and i differs by at most one. Matching is one-to-one before precision, recall, and F1 are computed. We first compute F1 for each RNA and then average within a task. Archive-II Macro is the unweighted mean of its nine task means.

B.4 Training Loss Definitions

For one RNA and update step t, let M be the nonempty set of valid upper-triangular candidate pairs, $y_{ij} \in \{0, 1\}$ the reference label, and

$$M_c = \{(i, j) \in M : y_{ij} = c\}.$$

Let

$$C = \{c \in \{0, 1\} : |M_c| > 0\}$$

contain the nonempty classes. The balanced BCE is

$$\mathcal{L}^{(t)}_{\text{bal-BCE}} = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|M_c|} \sum_{(i,j) \in M_c} \left[\text{softplus}(S^t_{ij}) - y_{ij} S^t_{ij}\right]. \tag{8}$$

Class balance uses separate positive and negative means, with empty classes omitted.

For the tolerance term, let τ denote the frozen backbone decision threshold used during training and set

$$q^{(t)}_{ij} = \sigma\left(S^t_{ij} - \text{logit}(\tau)\right)$$

with temperature 1. Before computing the tolerance neighborhood, q and y are set to zero outside the symmetric candidate mask $M \cup M^\top$. With zero padding at sequence boundaries, define

$$\mathcal{R}(A)_{ij} = \min\{1, A_{ij} + A_{i-1,j} + A_{i+1,j} + A_{i,j-1} + A_{i,j+1}\}. \tag{9}$$

The soft precision and recall for one RNA are

$$p_t = \frac{\sum_{(i,j) \in M} q^{(t)}_{ij} \mathcal{R}(y)_{ij}}{\max\left(\sum_{(i,j) \in M} q^{(t)}_{ij}, \epsilon\right)}, \qquad r_t = \frac{\sum_{(i,j) \in M} y_{ij} \mathcal{R}(q^{(t)})_{ij}}{\max\left(\sum_{(i,j) \in M} y_{ij}, \epsilon\right)}. \tag{10}$$

The tolerance loss is

$$\mathcal{L}_{\text{tol-F1}}^{(t)} = 1 - \frac{2p_t r_t}{\max(p_t + r_t, \epsilon)},$$

with $\epsilon = 10^{-8}$. Each loss is averaged over RNAs and then equally over the T updated states. The main configuration uses $T = 3$. The axial neighborhood permits a one-nucleotide shift of one endpoint but excludes simultaneous shifts of both endpoints. The soft surrogate does not impose the one-to-one assignment used by the discrete evaluation metric.

B.5 Sensitivity to Update Depth and BCE Weight

We assess sensitivity to update depth and BCE weight in the development/ablation setting. The sweep uses seed 42 on the Group I intron, tRNA, and tmRNA development tasks used in the main ablation. We report the unweighted mean flexible F1 across these three families. All other training and evaluation settings are held fixed.

Table S5 reports the complete development sweep. Mean family F1 is 0.7918, 0.8040, 0.8086, and 0.8083 for $T = 1, 2, 3, 4$ at $\lambda_{\text{BCE}} = 0.5$. Most of the improvement is obtained by the third update, and the fourth update provides no additional gain while requiring another application of the shared correction operator. We therefore use $T = 3$ in the main configuration.

At $T = 3$, the BCE-weight sweep gives 0.7702, 0.7949, 0.8086, and 0.8090 for $\lambda_{\text{BCE}} = 0, 0.25, 0.5, 1.0$. Performance is nearly unchanged between $\lambda_{\text{BCE}} = 0.5$ and 1.0, differing by 0.0004 mean F1. The main experiments use the fixed configuration $\lambda_{\text{BCE}} = 0.5$.

Table S5: **Development sensitivity of update depth and balanced-BCE weight.** Values are seed-42 macro flexible F1 over Group I intron, tRNA, and tmRNA. † marks the BCE weight used in the main experiments.

Sweep	Setting	Mean family F1
Update depth T	1	0.7918
	2	0.8040
	3	**0.8086**
	4	0.8083
λ_{BCE} ($T = 3$)	0	0.7702
	0.25	0.7949
	0.5†	0.8086
	1.0	0.8090

B.6 Structural Example Identifier

The 120-nt *Nicotiana tabacum* 5S rRNA shown in Figure 4 has identifier `5s_Nicotiana-tabacum-1`.

C Evidence Scope

The main evaluation reports family-held-out performance on nine Archive-II tasks. The ablation study uses the designated development families, and the structural analyses measure pair-level and loop-level changes from the corresponding model predictions. The supplementary material documents the deduplicated evaluation splits, model-selection procedure, architecture and optimization settings, thermodynamic configuration, perturbation protocol, decoding procedure, training objective, sensitivity analysis, and structural statistics referenced in the main text.