BASEPROMPT: CONTEXT AS PROMPT FOR CAUSAL GENOME LANGUAGE MODELS IN ZERO-SHOT RNA FITNESS PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal genome language models (GLMs) are pretrained on long genomic context, whereas zero-shot RNA fitness benchmarks often expose isolated mature transcripts with little usable local context. We study whether a frozen causal GLM can use synthetic context to improve prediction without task-specific training. BasePrompt is a training-free procedure that constructs short synthetic $5'$ and $3'$ flanks around a target sequence. Its key mechanism is reverse-complement (RC) decoding, which opens a left-context channel for an otherwise forward-only model: we decode on the RC strand to synthesize an upstream flank, map back to the original strand, and then decode a downstream flank autoregressively. The main evaluation uses a shared fixed budget of three bases per side and reports matched RC-ensemble and direct-scoring results separately. On RNAGym, this fixed-budget comparison improves all three Evo2 backbones in the main text; for example, Evo2 7B rises from 0.276 to 0.292 Spearman ρ under matched RC scoring. Gains persist without RC ensembling. Assay-level and type-level decompositions are used as descriptive robustness views, making the type-averaged benchmark aggregation and the remaining heterogeneity explicit. Auxiliary DNA controls show that extension content matters in the tested settings, without establishing the causal mechanism of the RNA gains. We present BasePrompt as an inference-time context-construction method for causal genome language models, not as a reconstruction of true biological flanks.

1 INTRODUCTION

Causal genome language models (GLMs) are typically pretrained on long genomic windows, yet downstream RNA benchmarks often expose only isolated mature transcripts rather than the surrounding genomic context seen during pretraining (Arora et al., 2025). This motivates a context-construction problem: the model is evaluated without much of the neighboring sequence available during pretraining. For decoder-only scorers, the asymmetry is especially sharp on the upstream side, because the model can naturally continue to the right but has no direct left-context generation channel. We test the utility of synthetic context in zero-shot RNA fitness evaluation, without assuming that context shortage causes the observed prediction errors.

Causal GLMs scale to long contexts, support likelihood-based scoring without task-specific heads, and are used as frozen backbones for mutation effect analysis and sequence generation (Nguyen et al., 2024; Merchant et al., 2024; Brixi et al., 2025; Wu et al., 2025). At the same time, bidirectional RNA encoders remain strong on several structured RNA tasks (Chen et al., 2022; Wang et al., 2024). We ask how to use these frozen causal models when an RNA benchmark provides little local context, rather than attempting a general comparison of genome and RNA model families.

We take an inference-time view of this problem. Instead of fine-tuning the backbone or introducing assay labels, we ask the model to construct a small amount of surrogate context in its native nucleotide alphabet (Figure 1). In this setting, the prompt is short sequence context rather than a natural-language instruction. The intervention has an explicit sequence budget and fits RNAGym, where each assay provides a reference sequence and many variants, making it natural to generate context once on the reference and then reuse it for the whole assay.

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.

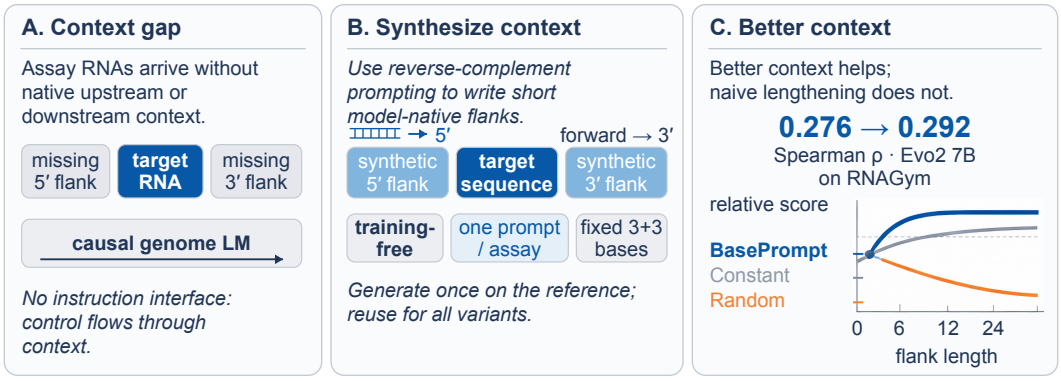

Figure 1: BasePrompt constructs context for causal genome language models. For isolated RNA inputs, a shared 3+3-base budget generates synthetic 5′ and 3′ flanks through reverse-complement and forward decoding, respectively. The same reference-generated flanks are reused for all variants, and the complete augmented sequences are scored. Under matched reverse-complement evaluation, Evo2 7B improves from 0.276 to 0.292 Spearman ρ. Panel C curves are schematic; the generated flanks are conditioning signals, not recovered biological flanks.

The key mechanism is that reverse-complement (RC) symmetry opens a left-context channel for an otherwise forward-only model. BasePrompt first maps RNA to the DNA alphabet by replacing U with T. It then decodes on the RC strand, maps the generated nucleotides back to the original strand to obtain a synthetic 5′ flank, and finally continues autoregressively on the original strand to generate a synthetic 3′ flank. Operationally, the prompt is generated once per assay reference sequence and reused for all variants in that assay, so flank generation scales with the number of references; scoring still processes each augmented mutant sequence individually with the same frozen GLM.

The generated flanks are synthetic conditioning signals, not recovered biological flanks. We do not infer transcript coordinates or claim that the model reconstructs the true genomic neighborhood of the assayed RNA. We test whether frozen causal GLMs benefit from this construction at inference time. We use a shared fixed budget of three bases per side for the main-text comparison and apply matched reverse-complement scoring to both the baseline and BasePrompt. Under this protocol, all three Evo2 backbones in the fixed-budget main comparison improve at the aggregate RNAGym level; gains remain visible without RC ensembling. Assay-level and type-level decompositions are used as descriptive robustness views, while auxiliary DNA controls compare extension content at matched budgets without identifying a causal mechanism for the RNA gains.

We summarize our contributions as follows:

- We introduce BasePrompt, a concrete *inference-time context-construction* procedure that synthesizes short flanking context in the model's native alphabet.
- We show how reverse-complement symmetry can be used at inference time to recover a left-context channel for a forward-only GLM, yielding a simple bidirectional context-generation procedure without architecture changes or task-specific supervision.
- Under a shared 3+3-base RNAGym protocol with matched RC scoring, we improve the three Evo2 backbones in the main comparison, while type-level and assay-level analyses make the heterogeneity across assays and RNA categories, and the scope of these results, explicit.

2 Related Work

Genome language models and zero-shot mutation scoring. Large-scale DNA and RNA foundation models support zero-shot prediction for mutation effect analysis, pathogenicity scoring, and sequence generation (Meier et al., 2021; Benegas et al., 2023; Zvyagin et al., 2023; Merchant et al., 2024; Brixi et al., 2025; Wu et al., 2025). GROVER learns contextual DNA representations, while Nucleotide Transformer evaluates pretrained representations across genomic prediction tasks and variant prioritization (Sanabria et al., 2024; Dalla-Torre et al., 2025). These studies connect se-

quence context to transferable genomic features. Causal GLMs are especially attractive when one wants likelihood-based scores from frozen backbones over long contexts. RNAGym provides strong zero-shot baselines for RNA fitness prediction (Arora et al., 2025). For causal decoders, this motivates investigating how the available sequence context is presented at evaluation time. Our work targets that inference-time presentation problem without retraining the backbone or adding labeled task data for parameter adaptation to individual assays.

RNA foundation models and RNA fitness evaluation. Bidirectional RNA models such as RNA-FM and RNAErnie have shown strong performance on structured RNA representation learning and downstream prediction (Chen et al., 2022; Wang et al., 2024). RiNALMo further examines transfer to structure prediction for unseen RNA families (Penić et al., 2025), while RNAGym standardizes fitness and structure evaluation across heterogeneous RNA datasets (Arora et al., 2025). Those models and the RNAGym benchmark clarify that RNA fitness prediction is not exclusively a DNA-LM problem. Our paper studies a narrower question: how far inference-time context construction can help *causal genome language models* on a zero-shot RNA fitness testbed. In other words, the contribution is an algorithmic intervention on the use of frozen causal GLMs, not a new RNA-pretrained architecture or a task-specific model trained on variant fitness labels.

Reverse-complement symmetry in genomics. Reverse-complement symmetry has been exploited in genomics through data augmentation, symmetry-aware architectures, and equivariant models (Cao & Zhang, 2019; Shrikumar et al., 2017; Bartoszewicz et al., 2020; Brown & Lunter, 2019; Schiff et al., 2024). Mallet & Vert (2021) characterize RC-equivariant layers, while Zhou et al. (2022) compare equivariant architectures and predictions combined across strands, including combination after training. These studies address how architectures and prediction procedures respect strand symmetry. BasePrompt uses the same symmetry to construct context. We use RC at inference time to let a forward-only causal decoder synthesize upstream context by generating on the reverse-complement strand and mapping back to the original strand. This separates context generation from RC ensembling, which combines scores computed from the two orientations of a sequence.

Prompting, context construction, and test-time adaptation. Prompting, search, retrieval, and related test-time methods improve frozen models by changing the input or decoding process rather than the backbone parameters (Brown et al., 2020; Kojima et al., 2022; Bahng et al., 2022; Jia et al., 2022; Liu et al., 2024; Fung et al., 2025). Related approaches, including prefix-tuning and prompt tuning, learn continuous task-specific prompts with a frozen backbone (Li & Liang, 2021; Lester et al., 2021). Our discrete nucleotide prompts are generated by the backbone itself, without optimizing additional prompt parameters. In biological sequence modeling, however, natural-language instructions are not the native control interface, and recent studies suggest that generic in-context learning recipes do not transfer cleanly to genomic models (Kantroo et al., 2025; Benegas et al., 2025b). BasePrompt sits closest to *context construction*: it changes what surrounding nucleotides the frozen GLM sees at evaluation time, using short synthetic sequence context written in the nucleotide alphabet accepted by the frozen backbone at inference time.

3 METHOD

BasePrompt is an inference-time context-construction algorithm for causal genome language models. Given an assay reference sequence, it spends a small prompt budget to synthesize short flanks in the model's own nucleotide alphabet. Figure 2 gives an overview, and Algorithm 1 summarizes the generation procedure for synthesizing both upstream and downstream flanks.

3.1 PROBLEM SETUP

Let r denote an RNA reference sequence for one assay, and let the variants in that assay be scored relative to r. We first convert r to a DNA string x by replacing U with T. This is only an alphabet conversion used to query DNA-trained GLMs; we do *not* claim to recover genomic coordinates or true biological flanks for the transcript. Let G be a causal genome language model, and let

$$\textsc{Generate}(G, s, L) \tag{1}$$

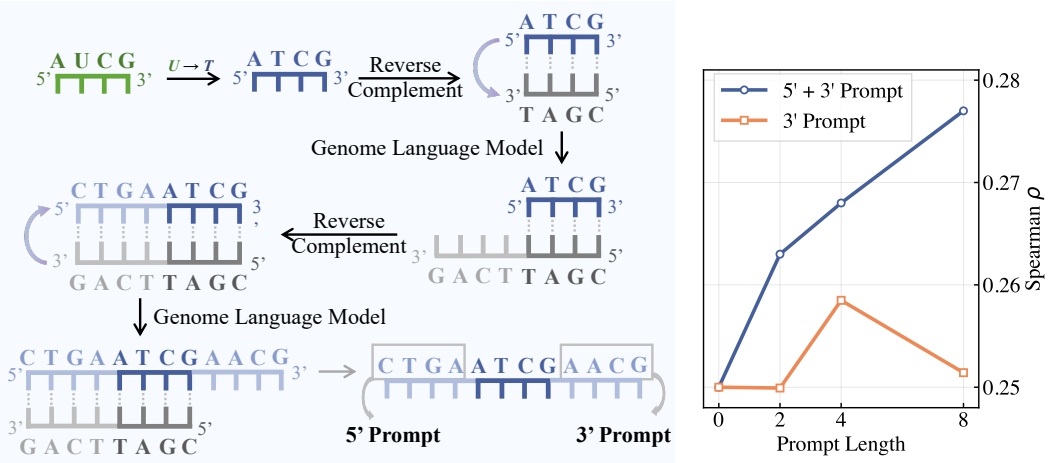

Figure 2: Overview of BasePrompt. *Left:* a causal GLM synthesizes an upstream flank through reverse-complement decoding and mapping back to the original strand, then generates a downstream flank autoregressively. *Right:* the displayed RNAGym comparison scores higher with both flanks than with the downstream flank alone. This observation motivates including upstream context without establishing why it improves prediction in this comparison.

return the L newly decoded nucleotides produced by G when conditioned on prefix s. The main RNAGym experiments use greedy decoding with outputs restricted to the DNA alphabet.

We also define a reverse-complement operator $\mathrm{RC}(\cdot)$ on nucleotide strings. For single-base tokenizers, RC directly reverses the sequence and swaps A↔T and C↔G. For k-mer tokenizers, we apply RC on the detokenized nucleotide string and retokenize afterward. This keeps the operation well defined at the nucleotide-string level and avoids ambiguous probability-level notation.

3.2 Constructing a bidirectional synthetic context

A causal GLM can naturally generate a downstream 3′ flank, but it has no direct mechanism for generating a left-context 5′ flank on the original strand. BasePrompt recovers that missing channel through RC symmetry. Given the DNAized input x, we first decode on the reverse-complement strand and then map the generated string back:

$$u_{1:L} = \mathrm{RC}(\textsc{Generate}(G, \mathrm{RC}(x), L)) \,. \tag{2}$$

This yields an upstream synthetic flank $u_{1:L}$ on the original strand. We then generate the downstream flank in the usual autoregressive direction,

$$d_{1:L} = \textsc{Generate}(G, u_{1:L} \circ x, L), \tag{3}$$

where $\circ$ denotes string concatenation. The final prompted input is

$$x^{+} = u_{1:L} \circ x \circ d_{1:L}. \tag{4}$$

Two implementation details matter. First, the 3′ flank is generated *after* the 5′ flank so that downstream decoding conditions on the same recovered left context later used for scoring. Second, prompting is performed only once per assay reference sequence, not once per mutant variant. Flank-generation work therefore scales with assay references, while scoring still processes every augmented variant under the chosen forward or RC-ensembled scoring protocol.

3.3 Variant scoring, RC fairness, and headline aggregation

For zero-shot variant scoring, every mutant in one assay inherits the same synthetic flanks generated from that assay's reference sequence. Let x^{+}_{ref} and x^{+}_{var} denote the prompted reference and prompted variant strings. We score a sequence by its mean token log-likelihood under the frozen GLM,

$$s_G(x) = \frac{1}{|x|} \sum_{t=1}^{|x|} \log p_G(x_t \mid x_{<t}), \tag{5}$$

Algorithm 1 BasePrompt with a fixed prompt budget L

Input: RNA reference sequence r, causal GLM G, prompt length L
Output: Synthetic flanks $u_{1:L}$ and $d_{1:L}$, prompted sequence x^+

$x \leftarrow \text{DNAIZE}(r)$ {replace U with T}
$x^{\text{rc}} \leftarrow \text{RC}(x)$
$\tilde{u} \leftarrow \text{GENERATE}(G, x^{\text{rc}}, L)$ {decode on RC strand}
$u_{1:L} \leftarrow \text{RC}(\tilde{u})$ {map back to upstream flank}
$d_{1:L} \leftarrow \text{GENERATE}(G, u_{1:L} \circ x, L)$
$x^+ \leftarrow u_{1:L} \circ x \circ d_{1:L}$
return $u_{1:L}, d_{1:L}, x^+$

and form the variant-effect score by comparing the prompted reference and variant:

$$\Delta_G(r, v) = s_G(x^+_{\text{ref}}) - s_G(x^+_{\text{var}}). \tag{6}$$

Higher values indicate that the mutant is less compatible with the pretrained sequence distribution than the reference. Both scores cover the complete augmented sequence, including the shared flanks. In the forward causal pass, the 3′ flank cannot alter likelihood terms at earlier RNA positions, but its own terms contribute to the score and depend on the preceding reference or variant sequence. This explains the computation, not why benchmark performance improves.

Following the Evo literature and RNAGym, we also consider reverse-complement (RC) ensembling at scoring time. Because RC averaging has a non-trivial effect on the reported numbers, the main comparison uses *matched RC scoring*: if the prompted model is evaluated with RC averaging, then the corresponding no-prompt baseline is evaluated with the same RC rule. We also report a no-RC direct-scoring control to isolate whether the gain survives without this averaging step. RC generation and RC ensembling are separate operations: the direct-scoring control still constructs its upstream flank through RC decoding, but does not average strand-specific scores.

RNAGym's headline benchmark score is *not* the plain mean across all 70 assays. For a metric m, RNAGym first averages within each RNA type and then averages those type-level scores with equal weight over the set of RNA types $\mathcal{T}$, rather than weighting types by their numbers of assays:

$$M_{\text{RNAGym}}(m) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{|\mathcal{A}_t|} \sum_{a \in \mathcal{A}_t} m(a), \tag{7}$$

where $\mathcal{A}_t$ is the set of assays for type t. This weighting is important later when we compare benchmark-level gains to assay-level win/loss counts and assay-wise mean deltas.

3.4 Budget selection and scope

The main-text comparisons use a shared budget of L=3 bases on each side for the evaluated Evo2 backbones. This provides a common comparison setting, not a universal optimum or a recommendation for every future task. Model-specific length sweeps are reported in the appendix as exploratory analyses, separately from the fixed-budget results.

Finally, the synthetic flanks are best viewed as *conditioning signals* for the downstream frozen GLM. They are not presented as recovered biological flanks, transcript-specific regulatory elements, or coordinate-resolved genomic reconstructions. The paper studies whether such synthetic context helps likelihood-based zero-shot evaluation, under tightly matched budgets and controls.

4 Experiments

4.1 Experimental Settings

We evaluate frozen GLMs without task-specific training. The same model constructs the synthetic context and performs the paired downstream predictions on the resulting inputs.

Table 1: Reference calibration for RC fairness on two backbones with source-compatible reported settings. Gray rows mark the settings reported by RNAGym and the matching re-implementation; the remaining rows show the score after switching the evaluation setting. This motivates matched RC protocols for baseline and BasePrompt throughout the fixed-budget Evo2 comparison.

Model	Setting	Spearman $\rho \uparrow$	AUROC $\uparrow$
Evo1.5	RNAGym reported	0.177	0.583
	Our re-impl. (direct)	0.177	0.583
	Our re-impl. (+ RC ensemble)	0.199	0.594
Evo2 7B	RNAGym reported	0.276	0.636
	Our re-impl. (direct)	0.251	0.622
	Our re-impl. (+ RC ensemble)	0.276	0.636

Benchmarks and metrics. RNAGym (Arora et al., 2025) contains 70 RNA mutational scanning assays and more than 1.1M variants across five RNA types. Spearman ρ, AUROC, and MCC use the type-averaged aggregation in Equation (7); assay-level summaries describe heterogeneity rather than statistical significance. Auxiliary ClinVar (Landrum et al., 2020; Benegas et al., 2025a) comparisons report absolute Spearman correlation and average precision under matched added-context budgets. The GENERator benchmark (Wu et al., 2025) measures base accuracy, absolute GC-content difference, and 2-mer Jensen–Shannon divergence. Its main comparison predicts six bases from a 192-base input with a 192-base upstream prompt. Dataset details are in Appendix A.

Models and protocol. The main RNA comparison covers Evo2 7B base, Evo2 7B, and Evo2 40B with matched-RC, fixed-budget direct, assay-level, and type-level results. Evo1 and Evo1.5 serve as auxiliary reference backbones (Nguyen et al., 2024; Merchant et al., 2024; Brixi et al., 2025). DNA comparisons also include Evo2 1B base and GENERator 1.2B/3B (Wu et al., 2025). RNAGym uses greedy prompting and three nucleotides per side, generated once per assay reference and reused for its variants. Each prompted/baseline pair shares its scoring protocol; direct scoring omits RC averaging. Evo uses single-nucleotide tokenization and GENERator uses 6-mers; auxiliary decoding details appear in Appendix A.4. Best-length sweeps and sampling checks remain secondary analyses, separate from the main fixed-budget comparison in the paper.

4.2 Main results on RNAGym

RC averaging can change benchmark scores, so Table 1 compares the reported settings and their re-implementations on two reference backbones. The subsequent fixed-budget comparisons match the scoring convention within each prompted/baseline pair and focus on the three Evo2 backbones.

BasePrompt improves the benchmark aggregate under the fixed 3+3 protocol. Under the shared fixed 3+3 setting, BasePrompt improves all three Evo2 backbones in the fully matched subset shown in Figure 3. The strongest result is Evo2 7B + BasePrompt with Spearman $\rho = 0.292$, improving over the matched Evo2 7B baseline (0.276). Because RNAGym reports a type-averaged benchmark score, we interpret these deltas as aggregate benchmark improvements rather than as per-assay significance claims or confidence estimates for the reported differences.

The gain is not solely an RC-ensemble artifact. Figure 4 reports direct prediction without RC averaging for the same three Evo2 backbones. BasePrompt improves each under the same 3+3 budget, so gains also occur without RC averaging. Model ordering changes with the scoring protocol: 40B exceeds 7B in direct scoring but not with RC ensembling. These comparisons do not identify model capacity as the cause of the protocol-dependent ordering of these checkpoints.

Assay-level views describe heterogeneity. Win/loss counts, median deltas, and assay-wise means in Figure 5 describe the distribution of improvements. They complement the type-averaged benchmark metric; they are neither confidence intervals nor per-assay significance tests. A win records the sign of a per-assay difference, so a majority of positive assays can coexist with a small or negative mean when positive and negative effects differ in magnitude across assays.

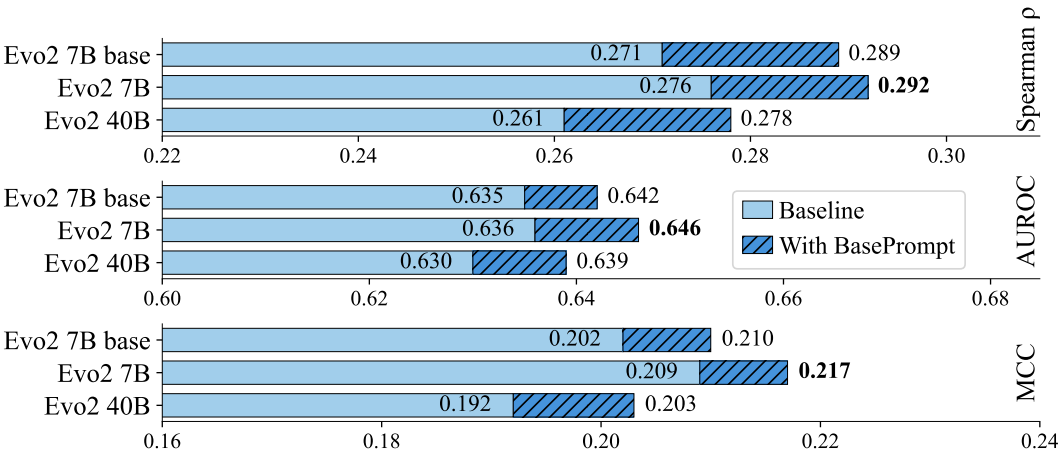

Figure 3: Main RNAGym results for Evo2 7B base, Evo2 7B, and Evo2 40B under the shared 3+3 comparison setting. Each model uses three synthetic bases per side, and both baseline and BasePrompt use matched reverse-complement scoring. Each metric is averaged within RNA type before overall aggregation. BasePrompt improves all three reported metrics for each backbone; Evo2 7B reaches the highest Spearman score, from 0.276 to 0.292. Assay-level heterogeneity is shown in Figure 5, with RNA-type scores in Appendix Table 8.

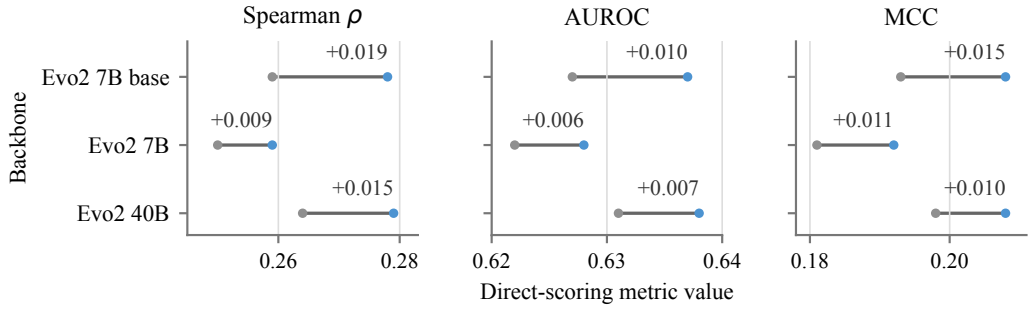

Figure 4: Direct-scoring control without reverse-complement ensembling for the same three Evo2 backbones used in the fixed-budget main comparison. Gray markers denote the no-prompt baseline and blue markers denote BasePrompt. BasePrompt improves Spearman ρ, AUROC, and MCC for each backbone under the same fixed 3+3 prompt budget. Exact values appear in Table 6.

Directional support is favorable for two backbones and mixed for one. Figure 5 summarizes fixed-budget matched-RC deltas for the three Evo2 backbones. Evo2 7B and Evo2 40B each improve on 41 of 70 assays, with positive median deltas of +0.0029 and +0.0016 and positive assay-wise mean deltas of +0.0030 and +0.0085. Evo2 7B base is the mixed case: it improves on 36 assays and declines on 34, its median delta is essentially zero (+0.0001), and its assay-wise mean delta is slightly negative (-0.0049), even though its benchmark-level aggregate still rises by +0.0175. The aggregate improvement therefore coexists with heterogeneous assay-level effects, especially for Evo2 7B base.

Type-level weighting explains the mixed aggregate/assay relation. The type-level breakdown makes the weighting issue concrete (Table 8). For Evo2 7B base, the prompted model gains strongly on mRNA splicing (+0.0478) and aptamer (+0.0565), stays near flat on tRNA and mRNA coding, and declines on ribozyme (-0.0269). Because RNAGym averages within RNA type before forming the overall score, these type-level gains can produce a positive headline aggregate even when the simple assay-wise mean delta is slightly negative. The assay-wise mean is therefore an auxiliary robustness view rather than the paper's target metric. The exact assay-level summary remains in Appendix Table 7 for comparison with the reported type-averaged scores.

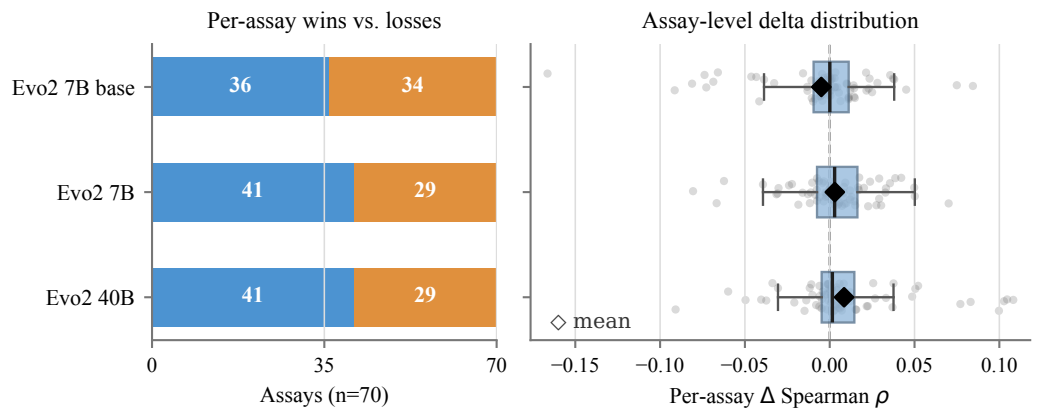

Figure 5: Assay-level RNAGym consistency under the fixed 3+3 matched RC protocol for the three Evo2 backbones. Left: per-assay win/loss counts relative to the matched no-prompt baseline; all three models have zero exact ties. Right: distribution of per-assay Spearman deltas, with boxplots over individual assays and black diamonds marking means. Evo2 7B and Evo2 40B are broadly favorable across assays, while Evo2 7B base remains mixed. The exact type-level breakdown of the matched baseline and prompted scores is reported in Appendix Table 8.

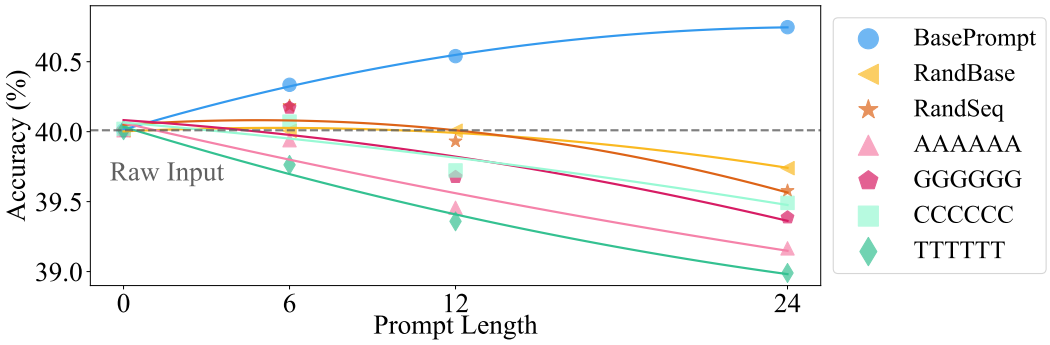

Figure 6: Context ablation on next-k-base prediction. Across the tested prompt budgets, BasePrompt improves as more synthetic context is added, while random and constant-base extensions do not show the same trend. The differing outcomes at matched lengths indicate that extension content matters in this DNA task; they do not isolate the mechanism of the RNA improvements.

Compute overhead is small in sequence length. For RNAGym, the main protocol generates only 6 extra nucleotides per assay reference and reuses them for all associated variants. Averaged over assay lengths, this corresponds to roughly 1.5% additional input length. This is *input-length overhead*, not a measured runtime, FLOP, or memory increase. Generation is reused across variants, but scoring costs still depend on the augmented length and model architecture.

4.3 Auxiliary DNA analyses

Same-budget native-flank evidence is supportive but secondary. ClinVar is useful because short native genomic flanks are available. Appendix Table 10 compares short native flanks and short synthetic flanks under identical extra-context budgets. Synthetic flanks outperform native flanks in 23 of 24 model/metric/budget cells. Evo2 1B Spearman at budget 6 is the exception: 0.6608 with synthetic flanks versus 0.6611 with native flanks. The table exposes the absolute score differences as well as their direction. We interpret this as same-budget efficiency for the downstream GLM, not as evidence that synthetic flanks are biologically superior.

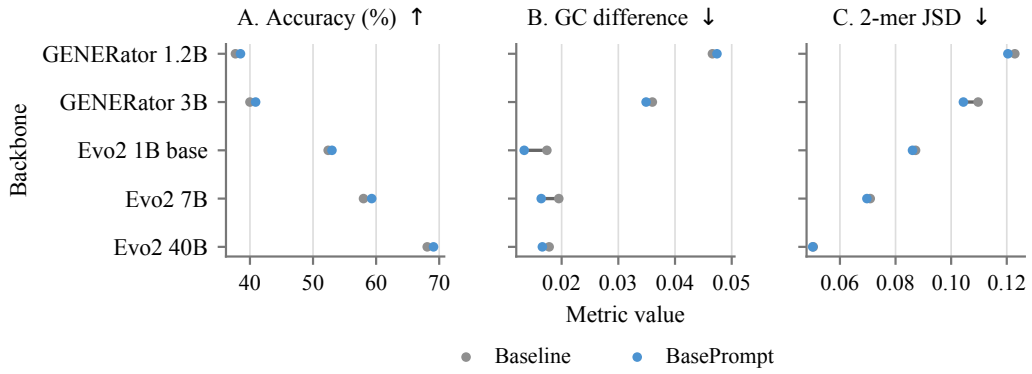

Figure 7: Next-6-base prediction with a 192-base input and a 192-base synthetic 5′ prompt. Each panel connects direct prediction and BasePrompt for one metric. Accuracy and 2-mer JSD improve on all five evaluated backbones. GC difference decreases on four backbones but increases for GENERator 1.2B. These sequence statistics provide a distributional check, rather than biological validation. Exact values appear in Appendix Table 11.

Context content matters in the tested DNA controls. Figure 6 compares BasePrompt to random-sequence extensions and fixed-base extensions. Only BasePrompt improves steadily as the prompt budget grows, whereas the alternative extensions do not show the same sustained improvement. This contrast supports the importance of extension content in these DNA settings; it does not isolate the causal mechanism responsible for the RNA fitness improvements in the main benchmark.

Synthetic flanks also improve short-horizon generation. On next-k-base prediction, BasePrompt improves base accuracy for every evaluated backbone and usually lowers GC-content difference and 2-mer Jensen–Shannon divergence relative to direct prediction in Figure 7. We treat this experiment as a distributional-quality sanity check rather than as direct biological validation of the prompts. Base accuracy measures position-wise agreement, whereas the other metrics describe coarser composition; matching composition does not imply reconstructing the target sequence. Exact metric values for the compared baselines and prompted models are reported in Appendix Table 11, separately for each evaluated backbone and each of the three generation metrics.

5 Conclusion

We presented BasePrompt, a training-free inference-time method that synthesizes short 5′ and 3′ flanking context for causal genome language models. The core idea is to spend a small prompt budget on context construction at inference time, using reverse-complement symmetry to recover a left-context channel for a forward-only model. The main fixed-budget evidence covers Evo2 7B base, Evo2 7B, and Evo2 40B. Under that uniform 3+3 protocol, the aggregate RNAGym score rises on all three, the gain persists without RC ensembling, and the descriptive assay-level views indicate favorable directional support for Evo2 7B and Evo2 40B and a benchmark-positive but heterogeneous pattern across assays and RNA categories for Evo2 7B base.

Limitations. Assay-level summaries describe heterogeneity but do not establish statistical significance or a causal explanation of the gains. We have not stratified RNA results by sequence length or tested truncating three bases at each end; the DNA controls do not replace these RNA analyses. Evidence for untested checkpoints remains limited, and GENERator is evaluated only on DNA tasks. Prompt utility also depends on the pretrained model and its tokenizer; these tests do not establish that the same budget or effects transfer to other backbones or biological settings. Synthetic flanks are conditioning signals, not recovered biological sequences.

AI USE STATEMENT

We used a GPT model to assist with drafting and rewriting portions of the manuscript text, including substantial English prose revisions for clarity, structure, and presentation. The model was used for writing assistance based on author-provided technical content, results, and outlines. All technical claims, numerical values, citations, and references were reviewed and verified by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies inference-time evaluation methods on public genomic benchmarks and does not involve human subjects, private data, or clinical deployment claims. The generated flanks are treated as synthetic conditioning signals for frozen models and are not presented as biological discoveries or medically actionable outputs for clinical decision-making.

REPRODUCIBILITY STATEMENT

The supplementary materials include the available experiment code, configurations, assay-level RNAGym results, and scripts with instructions for reproducing selected plots. The main text specifies the prompt construction, scoring, and aggregation rules; additional experimental settings and benchmark dataset descriptions are provided in Appendix A.

REFERENCES

Rohit Arora, Murphy Angelo, Christian Andrew Choe, Courtney A Shearer, Aaron W Kollasch, Fiona Qu, Ruben Weitzman, Artem Gazizov, Sarah Gurev, Erik Xie, et al. Rnagym: Large-scale benchmarks for rna fitness and structure prediction. *bioRxiv*, pp. 2025–06, 2025.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

Jakub M Bartoszewicz, Anja Seidel, Robert Rentzsch, and Bernhard Y Renard. Deepac: predicting pathogenic potential of novel dna with reverse-complement neural networks. *Bioinformatics*, 36 (1):81–89, 2020.

Gonzalo Benegas, Sanjit Singh Batra, and Yun S Song. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44): e2311219120, 2023.

Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, pp. 1–6, 2025a.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S Song. Genomic language models: opportunities and challenges. *Trends in Genetics*, 2025b.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, pp. 2025–02, 2025.

Richard C Brown and Gerton Lunter. An equivariant bayesian convolutional network predicts recombination hotspots and accurately resolves binding motifs. *Bioinformatics*, 35(13):2177–2184, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhen Cao and Shihua Zhang. Simple tricks of convolutional neural network architectures improve dna–protein binding prediction. *Bioinformatics*, 35(11):1837–1843, 2019.

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. *arXiv preprint arXiv:2204.00300*, 2022.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Sin-Hang Fung, Zhenghao Zhang, Ran Wang, Chen Miao, Brian Shing-Hei Wong, Kelly Yichen Li, Chenyang Hong, Jingying Zhou, Kevin Y Yip, Stephen Kwok-Wing Tsui, et al. Few-shot in-context learning with large language models for antibody characterization. *bioRxiv*, pp. 2025–02, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Pranav Kantroo, Günter P Wagner, and Benjamin B Machta. In-context learning can distort the relationship between sequence likelihoods and biological fitness. *arXiv preprint arXiv:2504.17068*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, et al. Clinvar: improvements to accessing data. *Nucleic acids research*, 48(D1):D835–D844, 2020.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Jiajia Liu, Mengyuan Yang, Yankai Yu, Haixia Xu, Kang Li, and Xiaobo Zhou. Large language models in bioinformatics: applications and perspectives. *arXiv preprint arXiv:2401.04155*, 2024.

Vincent Mallet and Jean-Philippe Vert. Reverse-complement equivariant networks for dna sequences. *Advances in neural information processing systems*, 34:13511–13523, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Aditi T Merchant, Samuel H King, Eric Nguyen, and Brian L Hie. Semantic mining of functional de novo genes from a genomic language model. *bioRxiv*, 2024. doi: 10.1101/2024.12.17.628962. URL https://www.biorxiv.org/content/early/2024/12/18/2024.12.17.628962.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://www.science.org/doi/abs/10.1126/science.ado9336.

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.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Reverse-complement parameter sharing improves deep learning models for genomics. *BioRxiv*, pp. 103663, 2017.

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.

Wei Wu, Qiuyi Li, Mingyang Li, Kun Fu, Fuli Feng, Jieping Ye, Hui Xiong, and Zheng Wang. Generator: A long-context generative genomic foundation model. *arXiv preprint arXiv:2502.07272*, 2025.

Hannah Zhou, Avanti Shrikumar, and Anshul Kundaje. Towards a better understanding of reverse-complement equivariance for deep learning models in genomics. In *Machine Learning in Computational Biology*, pp. 1–33. PMLR, 2022.

Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, et al. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 37(6):683–705, 2023.

Table 2: Information about our datasets. RNAGym serves as the main dataset for evaluation, while ClinVar and the next k-base prediction (NBP) are used as auxiliary datasets for analysis.

Dataset	Modality	Task	Species	Size
RNAGym	RNA	Fitness Prediction	Eukaryote, Prokaryote, Virus	1,117,995
ClinVar	DNA	Fitness Prediction	Human	40,976
NBP	DNA	Generation	Eukaryote, Prokaryote, Virus	84,648

APPENDIX

In the appendix, we first detail the experimental settings in Appendix A, provide an in-depth description of the BasePrompt mechanism in Appendix B, and present supplementary RNA and DNA analyses in Appendix C and Appendix D, respectively.

A EXPERIMENTAL SETTINGS

An overview of the evaluation tasks and datasets is provided in Table 2.

All experiments are conducted in a zero-shot setting, utilizing only the inference capabilities of pre-trained models without any fine-tuning. We evaluate BasePrompt on a compact set of RNA and DNA tasks that cover both discriminative and generative settings. The tasks are detailed in Appendices A.1 to A.3. We then describe the models and hyperparameters in Appendix A.4 and elaborate on the core zero-shot mechanism for variant effect prediction in Appendix A.5.

A.1 TASK-A: RNA VARIANT EFFECT PREDICTION ON RNAGYM

Task and Dataset We evaluate BasePrompt on the RNA variant effect prediction (VEP) task using the RNAGym benchmark (Arora et al., 2025). The objective is to predict the functional effect of RNA sequence variants in a zero-shot manner. Performance is evaluated by comparing the predicted functional ranking of variants against experimental measurements from deep mutational scanning (DMS) assays. The benchmark comprises 70 standardized DMS assays, covering 1,117,995 variants across diverse RNA types (*e.g.*, mRNA-splicing, tRNA, aptamer, ribozyme) and species (eukaryote, prokaryote, virus, human), with sequence lengths ranging from 45 to 5,592 bases. Detailed assay-level metadata and sequence lengths for RNAGym are listed in Table 3.

Evaluation and Metrics The reported RNAGym comparisons use three metrics: Spearman's rank correlation, area under the receiver operating characteristic curve (AUROC), and Matthews correlation coefficient (MCC). Spearman correlation measures agreement between predicted scores and experimental rankings, while AUROC and MCC assess discrimination between functional and non-functional variants. To ensure a balanced evaluation across RNA types with varying numbers of assays, we first compute the average performance for each RNA type and then report the overall mean of the resulting per-type averages with equal weight assigned to each RNA category.

A.2 TASK-B: NEXT-BASE PREDICTION

Task and Dataset We assess the generative capabilities of models on a DNA next-base prediction task, using data from various species, including eukaryote, prokaryote, virus, and human (Wu et al., 2025; Sanabria et al., 2024). The task requires the model to predict the next N bases given an input DNA sequence. We use the dataset from (Wu et al., 2025)[1], filtered to retain only sequences containing A, G, C, and T, resulting in 84,648 sequences. Dataset statistics are provided in Table 4.

Evaluation and Metrics We evaluate generation quality using three metrics that capture different aspects of performance. (1) Accuracy: the percentage of correctly predicted bases, calculated as

[1] https://huggingface.co/datasets/GenerTeam/next-kmer-prediction

Table 3: Assay names, variant counts, and sequence lengths for the RNAGym benchmark (Arora et al., 2025), grouped by the corresponding RNA type for each assay.

RNA Type	Dataset	Sequence Number	Sequence Length
Aptamer (Target binding ability)	Tome_2014_GFP_aptamer	417	82
	Tome_2014_NELFE_aptamer	2652	70
Ribozyme (Cleavage and splicing)	Andreasson_2020_ribozyme	7343	74
	Beck_2022_ribozyme	21321	69
	Janzen_2022_fam1a1_ribozyme	1953	71
	Janzen_2022_fam1b1_ribozyme	1953	71
	Janzen_2022_fam21_ribozyme	1953	71
	Janzen_2022_fam22_ribozyme	1953	71
	Janzen_2022_fam31_ribozyme	1953	71
	Kobori_2015_ribozyme_j12	255	230
	Kobori_2015_ribozyme_p4	255	190
	Kobori_2015_ribozyme_tw	1023	77
	Kobori_2016_osa_ribozyme	10296	54
	Kobori_2018_ribozyme	16383	160
	McRae_2024_5tu_ribozyme	74942	152
	McRae_2024_t1_ribozyme	47503	135
	Peri_2022_ribozyme	16383	197
	Pitt_2010_ribozyme	186	87
	Roberts_2023_HDV_ribozyme	33930	87
	Roberts_2023_cepeb3_ribozyme	21321	69
	Roberts_2023_hh_ribozyme	9045	45
	Roberts_2023_hp_ribozyme	22578	71
	Roberts_2023_tw_ribozyme	10296	48
	Soo_2021_ribozyme	64019	425
	Zhang_2020_cpeb3_ribozyme	111417	81
	Zhang_2024_OR4K15_ribozyme	61393	140
	Zhang_2024_line1_full_ribozyme	69583	146
	Zhang_2024_line1_mini_ribozyme	149710	46
	A0A2Z5U3Z0_9INFA_Doud_2016	10715	1698
mRNA Coding (Coding mRNA fitness)	BCHB_CHLTE_Tsuboyama_2023_2KRU	1652	156
	BLAT_ECOLX_Firnberg_2014	4783	861
	BLAT_ECOLX_Jacquier_2013	989	861
	BRCA1_HUMAN_Findlay_2018	886	5592
	C6KNH7_9INFA_Lee_2018	10754	1701
	CALM1_HUMAN_Weile_2017	1813	450
	CAPSD_AAV2S_Sinai_2021	42328	2208
	CBS_HUMAN_Sun_2020	7217	1656
	CCDB_ECOLI_Adkar_2012	1176	306
	DLG4_RAT_McLaughlin_2012	1576	2175
	DOCK1_MOUSE_Tsuboyama_2023_2M0Y	3099	198
	ESTA_BACSU_Nutschel_2020	2172	639
	F7YBW8_MESOW_Ding_2023	7922	282
	GFP_AEQVI_Sarkisyan_2016	51714	717
	HECD1_HUMAN_Tsuboyama_2023_3DKM	6835	216
	IF1_ECOLI_Kelsic_2016	1367	219
	MLAC_ECOLI_MacRae_2023	4007	636
	MTHR_HUMAN_Weile_2021	12464	1971
	OBSCN_HUMAN_Tsuboyama_2023_1V1C	3421	195
	OXDA_RHOTO_Vanella_2023_activity	6396	1092
	P53_HUMAN_Kotler_2018	1048	1182
	PKN1_HUMAN_Tsuboyama_2023_1URF	1305	213
	POLG_DEN26_Suphatrakul_2023	16897	2700
	POLG_PESV_Tsuboyama_2023_2MXD	6287	159
	PSAE_PICP2_Tsuboyama_2023_1PSE	1616	204
	PTEN_HUMAN_Mighell_2018	7260	1212
	Q837P4_ENTFA_Meier_2023	697	1770
	R1AB_SARS2_Flynn_2022	5725	918
	RCRO_LAMBD_Tsuboyama_2023_1ORC	2401	189
	RDRP_I33A0_Li_2023	12003	2271
	RNC_ECOLI_Weeks_2023	4277	681
	SPIKE_SARS2_Starr_2020_binding	3802	3822
	SPTN1_CHICK_Tsuboyama_2023_1TUD	3485	180
	SR43C_ARATH_Tsuboyama_2023_2N88	1663	144
	SUMO1_HUMAN_Weile_2017	1700	306
	TAT_HV1BR_Fernandes_2016	1577	258
mRNA Splicing (Splicing ability)	Julien_2016_mRNA	189	63
	Ke_2017_mRNA	5559	51
tRNA (Stability and growth)	Domingo_2018_tRNA	4175	72
	Guy_2014_tRNA	25491	105
	Li_2016_tRNA	65536	72

$(N_{correct}/N) \times 100\%$. (2) GC Content Difference: the absolute difference between the GC content ratio of the generated bases and the ground-truth bases. (3) k-mer Jensen-Shannon Divergence (JSD):

Table 4: Composition of the dataset used for auxiliary DNA next-base prediction.

Subset	Taxonomy / Species Type	Size
Eukaryote	Archaea, Viruses, Bacteria, Eukaryota	29,944
Bacteria	Bacteria	30,801
Others	Mitochondrion, Archaea, Plasmid, Virus, Plastid	23,903

the JSD between the k-mer frequency distributions of the generated and ground-truth sequences. These metrics provide a multi-grained analysis: accuracy measures base-level precision, GC content difference reflects compositional distribution, and k-mer JSD captures local sequence statistics.

A.3 TASK-C: CLINICAL VARIANT EFFECT PREDICTION

We further evaluate BasePrompt on a discriminative task focused on predicting the pathogenicity of human clinical single-nucleotide variants (SNVs). We use the ClinVar dataset (Landrum et al., 2020) to assess generalization on clinically significant variations. The model is given a pair of sequences—one containing the SNV and a corresponding reference—and must classify the variant as pathogenic or benign in a zero-shot setting. We adopt the version from (Benegas et al., 2025a)[2], which contains 40,976 samples for this clinical variant classification task.

Evaluation and Metrics The reported ClinVar metrics are Spearman's rank correlation and average precision (AP). Spearman correlation measures the association between the model's predicted pathogenicity scores and the ground-truth binary labels; it does not measure a graded scale of disease severity. AP summarizes precision and recall with pathogenic variants as the positive class, providing a complementary view of the resulting variant rankings.

A.4 MODELS AND HYPERPARAMETERS

The main RNA comparison evaluates Evo2-7B-base, Evo2-7B, and Evo2-40B (Brixi et al., 2025); Evo1 and Evo1.5 are additional reference backbones in the calibration and appendix analyses. The auxiliary DNA experiments evaluate GENERator-1.2B, GENERator-3B (Wu et al., 2025), Evo2-1B-base, Evo2-7B, and, for next-base prediction, Evo2-40B. The GENERator results therefore concern DNA tasks rather than the main RNA benchmark. For generative tasks, we employ standard sampling strategies: temperature, top-k, and top-p. Our default settings are chosen based on the models' tokenization schemes. For GENERator models, which use 6-mer tokenization, we set temperature=1.0, top-k=4, and top-p=1.0. For Evo2 models, which use single-nucleotide tokenization, generation is restricted to the four valid DNA bases (A, T, C, G), and top-k=1 selects greedy decoding. Prompt-decoding robustness for the fixed-budget RNAGym setting is discussed in Appendix C.5, including the effects of alternative top-k and top-p settings.

BasePrompt constructs prompts at inference time without altering the underlying model architecture or its downstream task pipeline. The specific application of BasePrompt to each task is detailed in Appendix B. For consistency and fair comparison, the same GLM is used for both generating the bidirectional prompts and for performing the final downstream task. For instance, an experiment labeled Evo2 7B uses Evo2 7B for both prompting and the subsequent prediction.

A.5 ZERO-SHOT MECHANISM OF VARIANT EFFECT PREDICTION

This section details the likelihood-based zero-shot variant classification used for clinical SNVs. The model is not trained on labeled variant effect data for this task; its pretrained sequence likelihoods provide the prediction scores. The input to the model consists of a pair of sequences: a reference sequence (ref) and a variant sequence (var), differing by a single nucleotide variant (SNV).

Here, the reference sequence $\mathbf{x}^{\text{ref}}$ represents the wild-type (non-mutated) version of a genomic region, while $\mathbf{x}^{\text{var}}$ is an otherwise identical sequence that contains an SNV at a specific position. The

[2]Dataset link.

clinical task distinguishes variants annotated as pathogenic from those annotated as benign. Given a reference sequence $\mathbf{x}^{\text{ref}}$ and a variant sequence $\mathbf{x}^{\text{var}}$, we compute a log-likelihood score for each sequence by averaging the model's log-probabilities over all positions:

$$\log p(\mathbf{x}) = \frac{1}{L} \sum_{t=1}^{L} \log p(x_t \mid \mathbf{x}_{<t}) \tag{8}$$

Here, L is the sequence length, and $p(x_t \mid \mathbf{x}_{<t})$ is the probability assigned to the true nucleotide at position t under the model's autoregressive output. In practice, this is computed by applying a log-softmax over the model's output logits at each position and gathering the value corresponding to the ground-truth token. The sequence-score difference below compares the reference and variant under the same model, using the likelihoods assigned to their respective complete inputs:

$$\Delta\mathcal{L} = \log p(\mathbf{x}^{\text{ref}}) - \log p(\mathbf{x}^{\text{var}}) \tag{9}$$

This score serves as a proxy for mutation impact, with higher values indicating greater disruption under the model's learned distribution. The reported rank correlation and average precision compare these variant-effect scores with the pathogenic/benign labels in the clinical evaluation.

B Additional Implementation Notes

These implementation notes supplement the algorithm, notation, and scoring protocol in the main text, with details on prompt reuse, generation order, and tokenization.

Reference-only prompt generation and transfer. For each assay, we generate the synthetic $5'$ and $3'$ flanks once from the reference sequence and reuse the same flanks for all associated variants. The reference and every associated variant are scored with these shared flanks over the full concatenated sequence. This reuse amortizes generation across the assay and avoids introducing independently generated prompts for different variants.

Order of generation. We first decode on the reverse-complement strand to construct the synthetic $5'$ flank and only then decode the $3'$ flank on the original strand. Downstream generation therefore conditions on the same synthetic left context used during scoring. The $3'$ flank does not change the likelihood of preceding RNA tokens in a causal model, but its own likelihood terms contribute to the score computed over the complete augmented sequence.

Tokenizer handling. For single-nucleotide tokenizers, the reverse-complement operator is applied directly on the nucleotide string. For k-mer tokenizers, we detokenize to the nucleotide string, apply reverse-complement at the string level, and retokenize before generation and scoring. This avoids any ambiguity about applying complement operations at the probability-scalar level.

RNA-to-DNA conversion. The U→T map is used only as an alphabet conversion so that DNA-trained GLMs can score the sequence. It is not intended as a reconstruction of transcript origin, genomic coordinates, or true biological flanking regions.

C Supplemented RNA Experiments

The main text reports fixed-budget results for Evo2 7B base, Evo2 7B, and Evo2 40B, including matched-RC, direct-scoring, and assay-level comparisons. This appendix supplies their exact numeric summaries and secondary sweeps, with Evo1 and Evo1.5 included as additional reference backbones for the reported exploratory comparisons.

C.1 Performance at Best Prompt Lengths

In the exploratory sweep in Figure 8 and Table 5, prompt length varies by backbone. Both the baseline and BasePrompt use reverse-complement ensembling. The selected lengths per side are

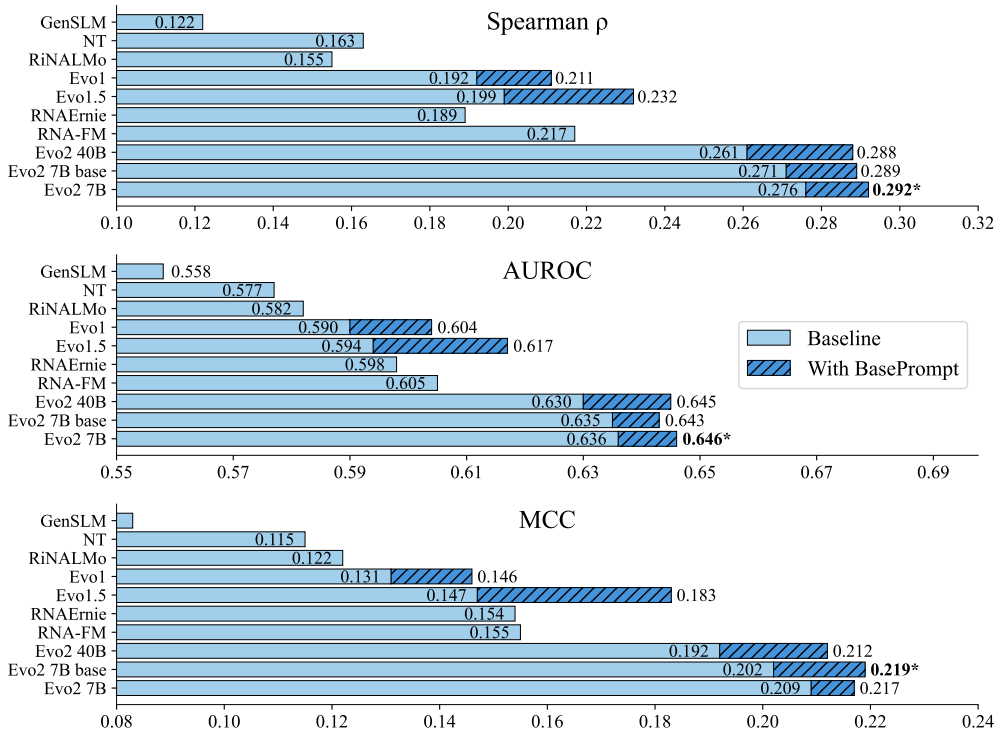

Figure 8: Exploratory RNAGym results with backbone-specific prompt lengths and matched reverse-complement ensembling for baseline and BasePrompt. Evo1 and Evo1.5 are included as additional reference backbones. These results are distinct from the shared 3+3 main comparison. Stars mark the highest displayed value for each metric, not statistical significance. Exact Evo scores and prompt lengths are reported in Table 5 and Appendix C.1.

32, 32, 5, 3, and 5 for Evo1, Evo1.5, Evo2 7B base, Evo2 7B, and Evo2 40B, respectively. These model-specific results are distinct from the shared 3+3 comparison setting, which is not a claim of a universally optimal length. The sweep does not include the full set of direct-scoring and assay-level comparisons reported for the fixed-budget main results.

C.2 Exact direct-scoring control

Table 6 provides the exact numbers visualized in the main-text direct-scoring control. The fixed 3+3 budget improves all three reported aggregate metrics for each of the three Evo2 backbones when RC averaging is removed from both baseline and BasePrompt. This comparison shows that improvements are also observed without RC averaging; it does not isolate the separate causal contributions of the upstream and downstream synthetic flanks to prediction performance.

C.3 Exact assay-level summary

Table 7 lists the win/loss counts across 70 assays and the mean/median deltas visualized in Figure 5. Evo2 7B and Evo2 40B have positive assay-wise means and medians, whereas Evo2 7B base has a slightly negative assay-wise mean despite its positive type-averaged aggregate. These descriptive summaries characterize heterogeneity without estimating statistical uncertainty.

C.4 RNA-type Breakdown

Table 8 provides the RNA-type scores behind the weighting discussion in the main text. Gains vary by category and backbone: splicing and aptamer scores improve in the displayed comparisons, coding changes are smaller, and some tRNA and ribozyme scores decline. Evo2 40B improves

Table 5: Exploratory RNAGym results with backbone-specific prompt lengths. Baseline and BasePrompt both use reverse-complement ensembling. Evo1 and Evo1.5 are additional reference backbones in this sweep; their results do not include the full set of direct-scoring and assay-level comparisons reported for the three Evo2 backbones in the fixed 3+3 main setting. Bold values mark the highest reported values of each metric among the models displayed in this table.

Model Name	Spearman $\rho \uparrow$	AUROC $\uparrow$	MCC $\uparrow$
RNAErnie	0.189	0.598	0.154
NT	0.163	0.577	0.115
RiNALMo	0.155	0.582	0.122
GenSLM	0.122	0.558	0.083
Evo1	0.192	0.590	0.131
w/ BasePrompt	0.211	0.604	0.146
Evo1.5	0.199	0.594	0.147
w/ BasePrompt	0.232	0.617	0.183
Evo2 7B base	0.271	0.635	0.202
w/ BasePrompt	0.289	0.643	**0.219**
Evo2 7B	0.276	0.636	0.209
w/ BasePrompt	**0.292**	**0.646**	0.217
Evo2 40B	0.261	0.630	0.192
w/ BasePrompt	0.288	0.645	0.212

Table 6: Direct-scoring RNAGym control without reverse-complement ensembling for the same three Evo2 backbones used in the fixed-budget main comparison. Each backbone still improves at the aggregate benchmark level under the same fixed 3+3 protocol, so the main gain is not solely an artifact of RC averaging. Cells report baseline $\rightarrow$ BasePrompt with absolute Δ.

Model	Spearman $\rho \uparrow$	AUROC $\uparrow$	MCC $\uparrow$
Evo2 7B base	0.259 $\rightarrow$ 0.278 (**+0.019**)	0.627 $\rightarrow$ 0.637 (**+0.010**)	0.193 $\rightarrow$ 0.208 (**+0.015**)
Evo2 7B	0.250 $\rightarrow$ 0.259 (**+0.009**)	0.622 $\rightarrow$ 0.628 (**+0.006**)	0.181 $\rightarrow$ 0.192 (**+0.011**)
Evo2 40B	0.264 $\rightarrow$ 0.279 (**+0.015**)	0.631 $\rightarrow$ 0.638 (**+0.007**)	0.198 $\rightarrow$ 0.208 (**+0.010**)

Table 7: Assay-level RNAGym summaries for the fixed 3+3 matched-RC protocol in Figure 5. Win/loss counts show how many of the 70 assays improve or decline. The overall column reports the change in the type-averaged benchmark aggregate and therefore differs from the simple assay-wise mean. These summaries are descriptive, rather than significance tests.

Model	Win	Loss	Median Δ assay	Mean Δ assay	Δ overall
Evo2 7B base	36	34	+0.0001	-0.0049	+0.0175
Evo2 7B	41	29	+0.0029	+0.0030	+0.0165
Evo2 40B	41	29	+0.0016	+0.0085	+0.0172

on four categories and ties its tRNA baseline at the reported precision. The ranking of Evo2 7B and Evo2 40B also varies with scoring protocol and RNA type: Evo2 7B has the higher matched-RC aggregate, while Evo2 40B has the higher direct-scoring aggregate. These comparisons do not establish a monotonic relationship between model scale and performance.

Table 8: RNA-type RNAGym scores under the fixed 3+3 budget and matched RC scoring. Among the displayed Evo-by-type comparisons, 21 of 25 deltas are non-negative. Gains vary by type and backbone: splicing and aptamer scores improve, coding changes are smaller, and some tRNA and ribozyme scores decline. Evo2 40B ties its tRNA baseline at the reported precision. Gray rows show the prompted scores obtained with BasePrompt for each corresponding backbone.

Model Name	mRNA-splic.	tRNA	Aptamer	Ribozyme	mRNA-cod.	Average
RNAErnie	0.230	0.416	0.037	0.142	0.117	0.189
NT	0.121	0.317	0.146	0.147	0.083	0.163
RiNALMo	0.348	0.260	0.026	0.072	0.070	0.155
GenSLM	0.173	0.093	0.126	0.135	0.084	0.122
Evo1	0.142	0.422	0.067	0.150	0.177	0.192
w/ BasePrompt	0.188	0.409	0.080	0.153	0.179	0.202
Evo1.5	0.146	0.430	0.082	0.164	0.173	0.199
w/ BasePrompt	0.179	0.424	0.085	0.164	0.173	0.205
Evo2 7B base	0.387	0.420	0.079	**0.186**	0.284	0.271
w/ BasePrompt	0.435	0.427	0.136	0.159	0.287	0.289
Evo2 7B	0.432	0.387	0.119	0.170	0.271	0.276
w/ BasePrompt	**0.468**	0.413	0.140	0.167	0.274	**0.292**
Evo2 40B	0.305	**0.431**	0.097	0.172	0.298	0.261
w/ BasePrompt	0.321	**0.431**	**0.152**	0.179	**0.305**	0.278

Table 9: Prompt-decoding sensitivity for Evo2 7B on RNAGym at the fixed 3+3 budget. The matched baseline is 0.276 / 0.636 / 0.209 for Spearman / AUROC / MCC. Several tested non-greedy settings improve all three metrics, while top-k=4 lowers MCC and top-p=0.9 lowers Spearman and ties the no-prompt baseline for MCC at the precision reported in this table.

Top-k	Top-p	Spearman ρ	AUROC	MCC
Baseline		0.276	0.636	0.209
1	-	0.292	0.646	0.217
2	-	0.296	0.648	0.223
3	-	0.295	0.648	0.221
4	-	0.280	0.641	0.207
-	0.3	0.282	0.641	0.211
-	0.5	0.296	0.648	0.222
-	0.7	0.282	0.641	0.211
-	0.9	0.272	0.637	0.209

C.5 Prompt-decoding robustness

At the fixed 3+3 budget, several tested top-k and top-p settings for Evo2 7B exceed the matched baseline (Table 9). The gains are not uniform: top-k=4 lowers MCC, while top-p=0.9 lowers Spearman and ties MCC at the reported precision. These results show sensitivity to decoding choices, alongside improvements under several non-greedy settings.

D Supplemented DNA Experiments

The auxiliary DNA analyses examine matched-budget native versus synthetic context, alternative sequence extensions, and short-horizon generation. This appendix reports the ClinVar and next-k-base values that accompany the main-text discussion. These DNA results do not establish the cause of the RNA improvements or replace RNA length-stratified and truncation controls.

Table 10: Matched-budget ClinVar comparison between native and synthetic flanks. Synthetic flanks outperform native flanks in 23 of 24 model/metric/budget cells under the same extra-context budget. Length 0 is the no-extra-context baseline. We interpret this as same-budget efficiency for the downstream GLM, not as evidence that synthetic flanks are biologically superior.

Spearman $\rho \uparrow$							
Model	0	Native 6	Ours 6	Native 12	Ours 12	Native 24	Ours 24
GENERator 1.2B	0.4871	0.4879	0.4907	0.4907	0.4924	0.4935	0.4980
GENERator 3B	0.5563	0.5575	0.5591	0.5587	0.5599	0.5605	0.5634
Evo2 1B	0.6595	0.6611	0.6608	0.6616	0.6622	0.6627	0.6635
Evo2 7B	0.6916	0.6921	0.6927	0.6928	0.6935	0.6936	0.6951
Average Precision ↑							
Model	0	Native 6	Ours 6	Native 12	Ours 12	Native 24	Ours 24
GENERator 1.2B	0.8304	0.8303	0.8313	0.8311	0.8320	0.8329	0.8342
GENERator 3B	0.8549	0.8552	0.8557	0.8559	0.8562	0.8571	0.8582
Evo2 1B	0.9070	0.9073	0.9078	0.9076	0.9083	0.9080	0.9087
Evo2 7B	0.9048	0.9052	0.9056	0.9056	0.9057	0.9057	0.9066

Table 11: Exact next-6-base results for the main-text distributional comparison, using a 192-base input and a 192-base synthetic 5′ prompt (Wu et al., 2025). Accuracy and 2-mer JSD improve on all five evaluated backbones; GC difference improves on four and worsens for GENERator 1.2B. These metrics describe local sequence statistics and do not establish biological realism.

Model	Accuracy ↑		GC Difference ↓		JSD (2-mer) ↓	
	Baseline	Ours	Baseline	Ours	Baseline	Ours
GENERator 1.2B	37.7	38.5	0.0466	0.0474	0.1229	0.1204
GENERator 3B	40.0	40.9	0.0360	0.0349	0.1097	0.1044
Evo2 1B base	52.4	53.0	0.0174	0.0134	0.0872	0.0861
Evo2 7B	58.0	59.3	0.0195	0.0164	0.0709	0.0697
Evo2 40B	68.1	69.1	0.0178	0.0166	0.0504	0.0502

D.1 Absolute ClinVar Table

The matched-budget comparison above is summarized in the main text. Synthetic flanks score higher than native flanks in 23 of the 24 reported model/metric/budget cells; the exception is Evo2 1B Spearman at budget 6. This is an observation about prediction under the tested short-context budgets, rather than evidence that synthetic flanks recover or improve on biological context.

D.2 Exact next-base summary

Table 11 gives the values visualized in the main-text next-k-base figure. Accuracy and 2-mer JSD improve on all five evaluated backbones. GC-content difference decreases on four backbones but increases for GENERator 1.2B, from 0.0466 to 0.0474. These local sequence statistics provide a distributional check, without establishing biological realism or a causal explanation for the gains.