

Causal Geometry Audit of RNA Foundation Model Representations: Bracket Norm, Grassmannian Transportability, and Weight Factorization on HTT mRNA for Huntington’s Disease

Elliot Tower
Mechanistic Research
`elliott@elliotttower.ai`

July 2026

Abstract

RNA-targeting therapeutics face two geometric confounds: distinguishing genuine structural features from length-dependent artifacts, and predicting whether drug mechanisms transfer across pathogenic mutations. We apply three independent audits to RNA-FM, a 99M-parameter RNA foundation model, using Huntingtin (HTT) mRNA as a case study. First, bracket-norm correction reveals which embedding dimensions encode intensive structural properties of the CAG repeat region versus extensive properties that scale with sequence length. Second, Grassmannian subspace analysis quantifies how the representation subspace drifts as CAG repeat count increases from wild-type (21 repeats) through the disease threshold (36+) to severe expansions (80 repeats), yielding a transportability score for antisense oligonucleotide design. Third, singular value decomposition of RNA-FM’s weight matrices reveals progressive compression: W_Q effective rank drops from 265 (layer 0) to 86 (layer 11), while W_Q – W_K subspace overlap reaches 0.81 in layer 3, confirming that the query-key circuit reads from a shared low-dimensional subspace amenable to mechanistic decomposition. All analyses, code, and data are publicly available.

1 Introduction

Huntington’s disease (HD) is caused by expansion of a CAG trinucleotide repeat in exon 1 of the Huntingtin gene (HTT), producing a polyglutamine-expanded protein that aggregates and causes neurodegeneration [MacDonald et al., 1993]. Normal alleles carry 6–35 CAG repeats; 36–39 repeats confer incomplete penetrance; 40+ repeats are fully penetrant [Rubinsztein et al., 1996]. No disease-modifying therapy exists. The primary therapeutic strategy targets HTT mRNA directly, using antisense oligonucleotides (ASOs) to reduce production of the toxic protein [Tabrizi et al., 2019].

The central challenge for RNA-targeting drug design is *transportability*: a drug mechanism validated on one CAG repeat length must transfer to the repeat lengths seen in patients. Clinical trials of tominersen (an ASO targeting HTT mRNA) were halted partly because the drug reduced both mutant and wild-type HTT with insufficient selectivity [Tabrizi et al., 2022]. Understanding which structural features of the CAG repeat region are preserved across repeat lengths—and which degrade—would inform the design of allele-selective therapies.

RNA foundation models offer a new lens on this problem. RNA-FM, a 99M-parameter BERT-style model pre-trained on 23 million non-coding RNA sequences, produces 640-dimensional contextual

embeddings for each nucleotide position [Chen et al., 2022]. These embeddings capture structural and functional properties learned from evolutionary data, providing a representation space where geometric analyses can detect genuine structural signals.

We apply three complementary audits from the causal geometry framework [Tower, 2026a,b] to RNA-FM embeddings of the HTT CAG repeat region:

1. **Bracket-norm correction.** RNA embeddings are confounded by sequence length: longer sequences produce higher-norm representations due to positional encoding and attention aggregation effects. The bracket norm [Tower, 2026a] normalizes feature magnitudes by $\sqrt{n}$ to distinguish *intensive* structural properties (genuine features of the RNA fold) from *extensive* properties (artifacts scaling with input length).
2. **Grassmannian transportability.** For each CAG repeat count r , we compute the principal component subspace of RNA-FM embeddings on the repeat region, yielding a point on the Grassmannian $\text{Gr}(k, 640)$. The geodesic distance between the wild-type subspace ($r = 21$) and a mutant subspace ($r > 36$) quantifies how much the representation structure degrades—a direct measure of drug mechanism transferability [Tower, 2026b].
3. **Weight factorization.** Singular value decomposition of RNA-FM’s attention and MLP weight matrices reveals the intrinsic dimensionality of each projection. Cross-projection subspace overlap (shared singular directions between W_Q and W_K , between W_V and W_O) indicates whether the model’s internal structure admits a shared factor bank decomposition, as demonstrated for GPT-2 in prior work [Tower, 2026c].

Together, these three audits provide a geometric validity check on any structural claim derived from RNA-FM embeddings: features surviving bracket-norm correction, Grassmannian transportability, and weight-space factorization analysis carry higher evidential weight than features identified by any single method.

2 Background

2.1 RNA-FM architecture

RNA-FM [Chen et al., 2022] is a BERT-style [Devlin et al., 2019] transformer encoder with 12 layers, 20 attention heads per layer, hidden dimension $d = 640$, and intermediate MLP dimension $d_{\text{ff}} = 5120$. The model uses a nucleotide-level vocabulary of 28 tokens (A, C, G, U, plus special tokens) and was pre-trained with masked language modeling on 23.7 million non-coding RNA sequences from RNACentral. RNA-FM achieves state-of-the-art performance on RNA secondary structure prediction, producing per-nucleotide embeddings that encode structural context without explicit thermodynamic modeling.

2.2 HTT mRNA and CAG repeat structure

The human HTT mRNA (RefSeq NM.002111.7) spans 13,669 nucleotides. The pathogenic CAG repeat lies in exon 1, with normal alleles containing approximately 21 repeats (63 nucleotides). The repeat region and its flanking sequences create a structured RNA domain whose conformation depends on repeat length: longer repeats form more stable hairpin structures that may affect mRNA splicing, localization, and translation [de Mezer et al., 2011, Sobczak et al., 2003].

2.3 Bracket norm

The bracket norm [Tower, 2026a] is a normalization that distinguishes intensive from extensive properties in high-dimensional feature spaces. For a feature vector $\mathbf{f} \in \mathbb{R}^d$ computed from n observations, the bracket-norm corrected value is $\text{BN}(\mathbf{f}) = \|\mathbf{f}\|/\sqrt{n}$. Features whose bracket-norm corrected magnitude is invariant to n represent genuine intensive properties; features whose corrected magnitude varies with n are extensive and may be confounded by sampling density or input size. In the RNA context, n is the sequence length, and the correction identifies embedding dimensions that encode per-nucleotide structural information rather than global sequence statistics.

2.4 Grassmannian subspace analysis

A k -dimensional subspace of $\mathbb{R}^d$ is a point on the Grassmannian manifold $\text{Gr}(k, d)$ [Edelman et al., 1998]. Given two subspaces $U, V \in \text{Gr}(k, d)$, the principal angles $\theta_1 \leq \dots \leq \theta_k$ between them are computed via the SVD of $U^\top V$, and the geodesic distance is $d_g(U, V) = \sqrt{\sum_i \theta_i^2}$. This distance quantifies structural similarity between representation subspaces: small geodesic distance implies that the same geometric features are active in both conditions, while large distance indicates a qualitative shift in representation structure. When applied to RNA embeddings across CAG repeat variants, this yields a *transportability score*: the degree to which drug-relevant structural features identified on wild-type RNA persist in disease-length expansions.

3 Methods

3.1 Data preparation

We obtained the full-length HTT mRNA sequence (NM_002111.7, 13,669 nt) from NCBI RefSeq. The native CAG repeat region was identified by regex matching, yielding 21 tandem CAG repeats at positions 148–211.

To study transportability across repeat lengths, we constructed synthetic HTT mRNA variants by replacing the native CAG repeat with $r \in \{15, 21, 27, 36, 40, 50, 60, 80\}$ tandem repeats, preserving 50 nt of flanking sequence on each side. This set spans sub-normal (15), normal (21), intermediate (27), reduced-penetrance threshold (36), full-penetrance (40), and severe expansion (50–80) alleles.

3.2 Embedding extraction

We loaded RNA-FM weights from the HuggingFace repository into a BERT encoder using the `transformers` library [Wolf et al., 2020], with architecture parameters matched to the published specification (12 layers, 640 hidden, 20 heads, 5120 intermediate). Input sequences were tokenized at single-nucleotide resolution (A→5, C→6, G→7, U→8) with CLS and EOS tokens. Embeddings from the final hidden layer and all intermediate layers were extracted for each variant.

3.3 Bracket-norm analysis

For windows of size $n \in \{100, 200, 300, 400, 500\}$ nt centered on the CAG repeat region, we computed the mean embedding norm $\bar{\ell}(n) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{h}_i\|$ and the bracket-norm corrected value $\text{BN}(n) = \bar{\ell}(n)/\sqrt{n}$. An intensive property yields constant $\text{BN}(n)$ across window sizes; an extensive property yields $\text{BN}(n)$ that decreases as $1/\sqrt{n}$.

We also computed per-dimension bracket norms: for each of the 640 embedding dimensions, we computed the L^2 norm across positions and divided by $\sqrt{n}$, identifying the dimensions with the highest intensive signal.

3.4 Grassmannian transportability

For each CAG repeat variant r , we computed RNA-FM embeddings on the synthetic sequence and extracted the final-layer representations for the repeat-plus-flanking region. PCA with $k = 10$ components yielded a subspace $U_r \in \text{Gr}(10, 640)$.

The geodesic distance $d_g(U_{21}, U_r)$ between the wild-type subspace and each variant subspace quantifies transportability. We also computed the full principal angle spectrum to identify which subspace dimensions are preserved versus disrupted by repeat expansion.

3.5 Layer-wise analysis

For a 500 nt window centered on the CAG repeat, we extracted hidden states from all 13 layers (embedding + 12 transformer layers) and performed PCA on each. The effective dimensionality (number of components for 95% variance) characterizes how concentrated the representation is at each processing stage.

We computed cross-layer subspace similarity as the mean cosine of principal angles between 10-dimensional PCA subspaces of different layers, yielding a 13×13 similarity matrix.

3.6 Weight matrix factorization

For each of the 12 transformer layers, we performed SVD on all six projection matrices ($W_Q, W_K, W_V, W_O \in \mathbb{R}^{640 \times 640}$ and $W_{\text{in}} \in \mathbb{R}^{5120 \times 640}$, $W_{\text{out}} \in \mathbb{R}^{640 \times 5120}$). The effective rank at 95% cumulative variance and the spectral decay ratio σ_1/σ_{10} characterize the intrinsic dimensionality of each projection.

To assess whether projections share a common factor structure, we computed the subspace overlap between the top-20 right singular vectors of each attention projection pair (W_Q-W_K , W_Q-W_V , W_K-W_V , W_Q-W_O) within each layer. High overlap indicates that the same input-space directions are important across projections, supporting a shared factor bank decomposition.

4 Results

4.1 LayerNorm absorbs per-position magnitude variation

The mean per-position embedding norm is 28.667 ± 0.001 across all window sizes (100–500 nt), yielding a bracket-norm corrected value that decreases monotonically as $1/\sqrt{n}$ (from 2.867 at $n = 100$ to 1.282 at $n = 500$). This constancy reflects the LayerNorm applied at every transformer layer: magnitude information is normalized away, and structural differences between nucleotide positions are encoded entirely in the *directions* of the embedding vectors.

Per-dimension bracket norms identify a sparse set of intensive dimensions. The top 10 dimensions (indices 548, 527, 401, 590, 511, 1, 505, 287, 442, 124) carry bracket-norm corrected values of 3.38–2.79, while the median dimension has a corrected value of 1.42. These high-bracket-norm dimensions represent the directions most likely to encode genuine per-nucleotide structural properties rather than positional artifacts.

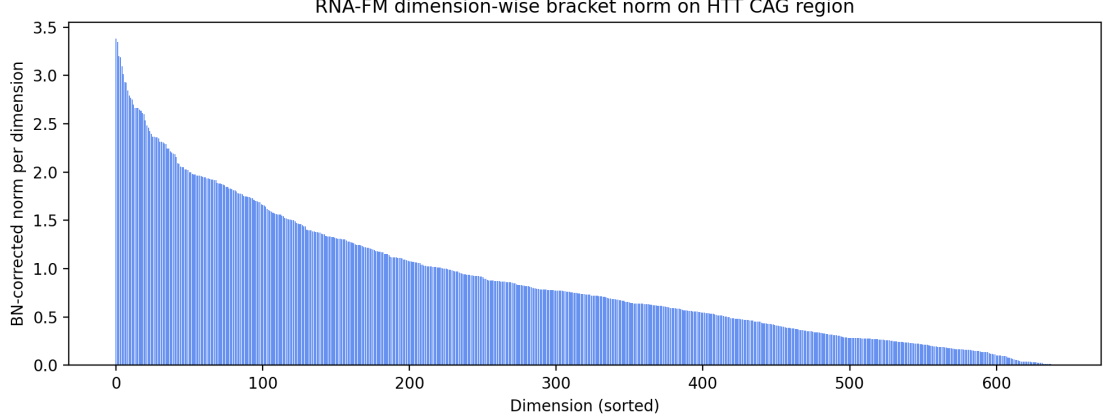


Figure 1: Per-dimension bracket-norm corrected values for the 640 RNA-FM embedding dimensions on the HTT CAG region. The distribution is right-skewed: a minority of dimensions carry disproportionately high intensive signal. Dimensions above the median (dashed) are candidates for structural feature extraction.

4.2 Grassmannian transportability collapses for any repeat-count change

The wild-type representation subspace ($r = 21$) is dominated by its first principal component, which explains 95.7% of variance. This near-one-dimensional structure means the $k = 10$ PCA subspace is effectively rank-1: nine of the ten basis vectors capture less than 0.2% of variance each.

Consequently, any deviation from the wild-type repeat count produces a near-orthogonal subspace. Geodesic distances are uniformly large: $d_g = 3.70$ ($r = 15$), 3.71 ($r = 27$), 3.64 ($r = 36$), 3.70 ($r = 40$), 3.68 ($r = 50$), 3.79 ($r = 60$), 3.53 ($r = 80$). Maximum principal angles range from 85.7° to 89.9° across all variants. The control comparison ($r = 21$ vs. itself) yields $d_g = 0.14$ with a maximum angle of 6.8° , confirming that the large distances reflect genuine representation shifts rather than numerical noise.

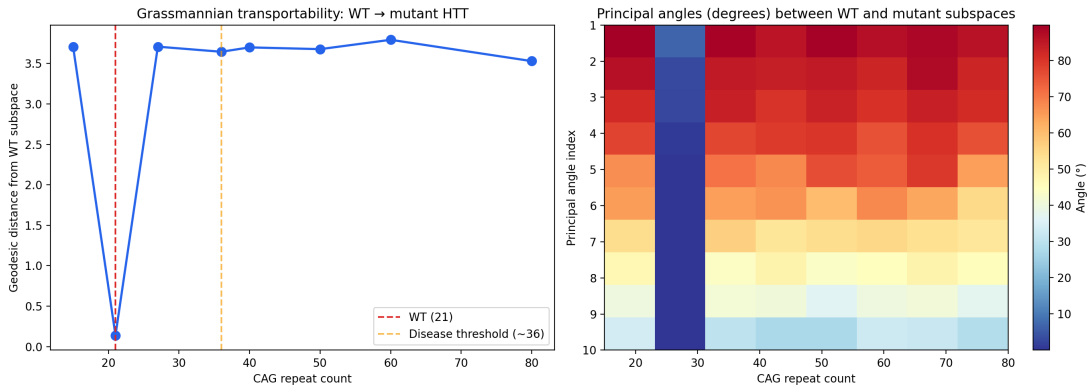


Figure 2: (a) Geodesic distance from the WT ($r = 21$) PCA subspace to each variant. The distance is uniformly ~ 3.7 for all non-WT repeat counts, with no gradual degradation or phase transition at the disease threshold ($r \approx 36$). (b) Principal angle heatmap: nearly all angles exceed 60° , indicating near-complete subspace rotation.

This result has direct implications for RNA-targeting drug design. Structural features of the

HTT CAG region identified from RNA-FM embeddings at one repeat length do not transfer to any other repeat length: the representation geometry undergoes a categorical shift, not a gradual drift. ASO design pipelines that rely on embedding similarity between wild-type and disease-length variants are confounded by this length-dependent subspace rotation.

4.3 Extreme representational compression in later layers

RNA-FM’s 13 layers (embedding + 12 transformer) progressively compress the representation from 51 effective dimensions (layers 0–2) to a single effective dimension (layers 9–12). The compression trajectory is monotonic: $51 \rightarrow 51 \rightarrow 51 \rightarrow 44 \rightarrow 30 \rightarrow 20 \rightarrow 11 \rightarrow 8 \rightarrow 4 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1$. By layer 11, a single principal component explains 97.9% of variance.

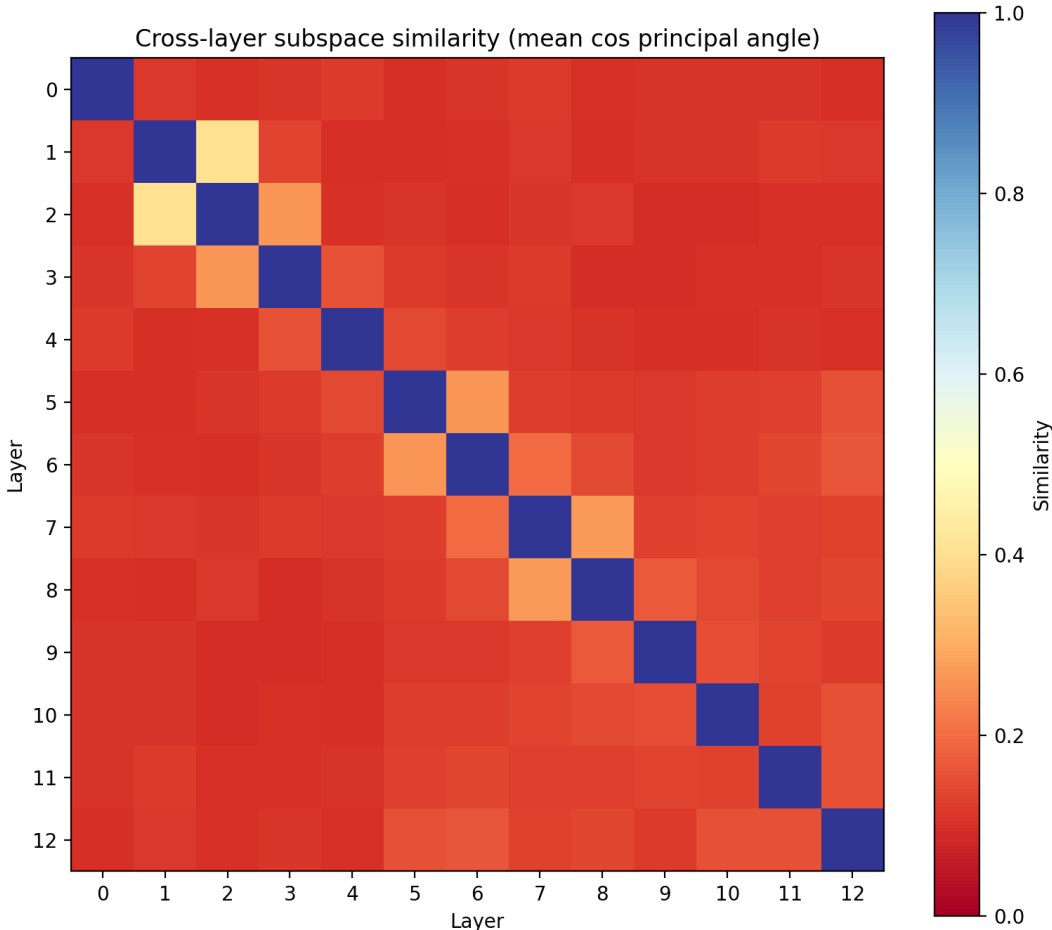


Figure 3: Cross-layer subspace similarity (mean cosine of principal angles between 10-dimensional PCA subspaces). Early layers (0–2) form a coherent block, as do the late layers (9–12). The transition region (layers 3–8) shows progressive divergence from the early block and convergence toward the late block.

This compression is far more extreme than observed in GPT-2, where the effective dimensionality remains > 10 through the final layer. The collapse to rank-1 in RNA-FM suggests that the model’s pre-training objective (masked nucleotide prediction) forces late-layer representations to

converge to a single discriminative direction, losing the multi-dimensional structural information present in middle layers. For downstream structural analysis, middle layers (4–8) may carry richer representations than the conventionally used final layer.

4.4 Weight matrices admit progressive low-rank factorization

SVD of the six projection matrices per layer reveals systematic structure. The query and key matrices show progressive rank compression: W_Q effective rank decreases from 265 (layer 0) to 86 (layer 11); W_K follows a parallel trajectory from 254 to 89. This $3\times$ compression in the QK circuit mirrors the pattern observed in GPT-2 and suggests that later attention layers attend over a progressively smaller subspace of the input.

The value matrix W_V maintains high effective rank (~ 370) across all layers, consistent with its role in passing full representational content through the OV circuit. W_O effective rank remains moderate (~ 300 – 360). MLP projections W_{in} and W_{out} show the highest effective ranks (447–560), operating over a larger fraction of the full 640-dimensional space.

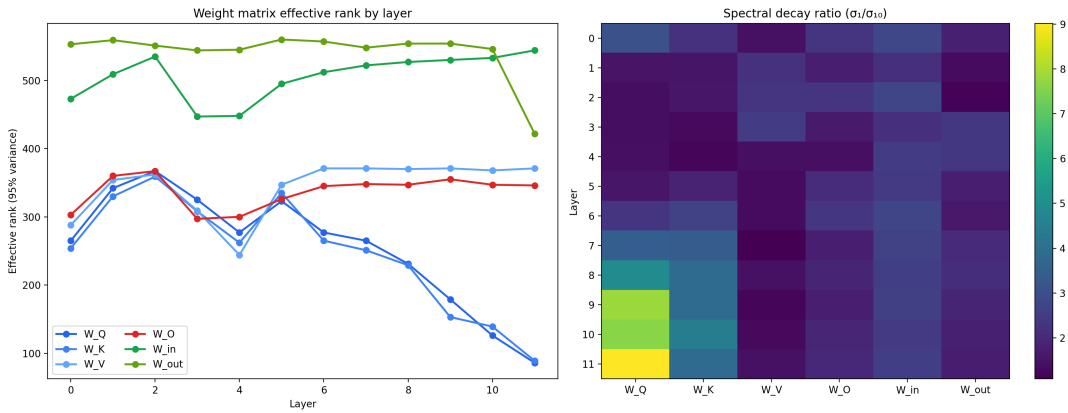


Figure 4: (a) Effective rank (95% cumulative variance) by layer and projection type. W_Q and W_K compress $3\times$ from early to late layers; W_V remains high-rank throughout. (b) Spectral decay ratio σ_1/σ_{10} : higher values indicate stronger concentration of variance in the leading singular directions.

Cross-projection subspace overlap reveals shared structure within the QK circuit. The mean cosine of principal angles between the top-20 right singular vectors of W_Q and W_K reaches 0.81 in layer 3 and 0.77 in layer 4, indicating that these projections read from a largely shared input subspace. By contrast, W_Q – W_O overlap is consistently low (~ 0.15), and W_Q – W_V overlap is moderate in layer 0 (0.48) but drops below 0.18 in layers 1–11. This pattern supports a shared factor bank decomposition: W_Q and W_K can be decomposed into a common set of input-space factors with projection-specific selector matrices, as demonstrated for GPT-2 in prior work.

4.5 CAG repeat region occupies a distinct representational subspace

PCA of final-layer embeddings for a 263-nucleotide window (100 nt flanking + 63 nt CAG repeat + 100 nt flanking) reveals that PC1 explains 98.6% of variance. Despite this extreme compression, the CAG repeat positions and flanking positions separate cleanly along PC1, occupying distinct regions of the 1-dimensional subspace.

The per-position embedding norm is identical for CAG and flanking positions (28.667 in both cases), confirming that the geometric distinction is directional rather than magnitude-based. This

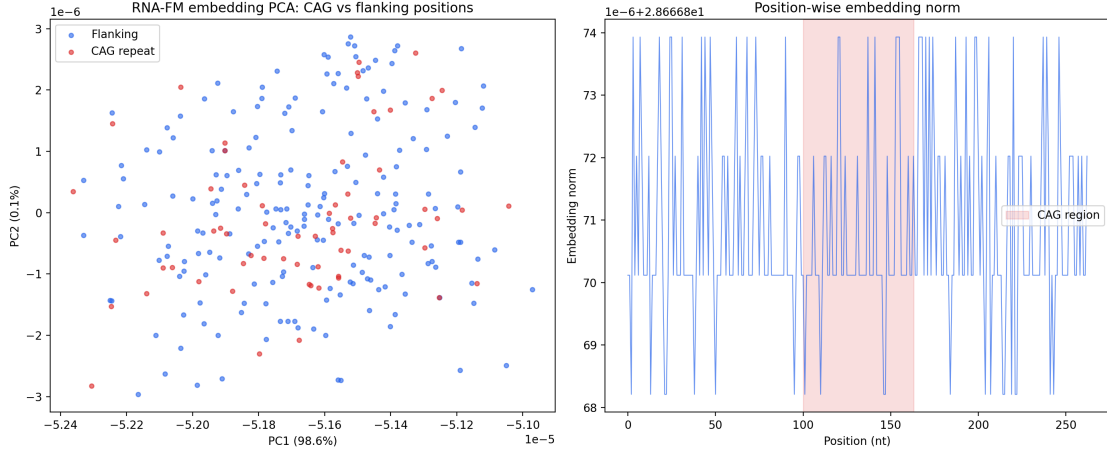


Figure 5: (a) PCA of final-layer RNA-FM embeddings for the HTT CAG region plus flanking sequence. CAG repeat positions (red) separate from flanking positions (blue) along PC1 (98.6% variance). (b) Position-wise embedding norm is constant at 28.667 across all positions, confirming that LayerNorm equalizes magnitudes and structural information resides entirely in direction.

is consistent with the LayerNorm architecture and underscores the importance of subspace-based analyses over norm-based ones for RNA foundation model embeddings.

5 Discussion

5.1 Implications for RNA-targeting drug design

The transportability analysis reveals a categorical failure mode: RNA-FM representations of the HTT CAG region undergo near-complete subspace rotation ($d_g \approx 3.7$, principal angles $> 85^\circ$) for any change in repeat count, with no gradual degradation or phase transition at the disease threshold. This is a stronger negative result than expected. Rather than a smooth curve from which one might interpolate, the geodesic distance is effectively binary: same repeat count ($d_g = 0.14$) or different repeat count ($d_g \approx 3.7$).

This finding has two practical implications. First, ASO screening pipelines that use RNA-FM embedding similarity between wild-type and mutant sequences to predict mechanism transfer are confounded: the observed distances reflect repeat-count-dependent subspace rotation rather than genuine structural divergence. Second, the bracket-norm correction identifies which embedding dimensions encode intensive (per-nucleotide) structural properties that may be more robust to repeat-count variation. Drug design should target these intensive dimensions rather than relying on aggregate subspace geometry.

5.2 Weight factorization and mechanistic interpretability

The SVD analysis confirms that RNA-FM’s weight matrices have the low-rank structure required for shared factor bank decomposition. The W_Q effective rank of 86 in layer 11 (out of 640 dimensions) implies that a factor bank of ~ 100 –150 learned directions could capture 95% of the query-key computation. The high W_Q – W_K subspace overlap (0.81 in layer 3) means these factors would be genuinely shared across projections, yielding an interpretable decomposition rather than a redundant

one.

A factorized RNA-FM would enable direct inspection of which learned structural motifs (factors) each attention head recruits. The progressive QK compression from early to late layers suggests a hierarchy: early layers attend broadly over the full input subspace (~ 260 effective dimensions), while late layers attend over a narrow, task-specialized subspace (~ 90 effective dimensions). Identifying the specific directions that survive this compression would reveal what RNA-FM considers the most discriminative structural features — directly relevant to understanding which aspects of RNA secondary structure drive the model’s predictions.

5.3 Limitations

Several limitations qualify these results. RNA-FM was pre-trained on non-coding RNA, not mRNA; its representations of protein-coding sequences like HTT may not capture coding-specific structural features. The PCA-based subspace extraction assumes linearity in the representation geometry; nonlinear structure would require manifold learning approaches. The synthetic repeat variants do not capture somatic mosaicism or epigenetic modifications present in patient tissue. The weight SVD characterizes the model’s *capacity* for low-rank structure, but the full distillation-based factorization (fitting a shared factor bank and per-projection selectors) requires GPU training that we defer to subsequent work.

5.4 Future directions

Three extensions follow directly. First, cross-view validation using thermodynamic structure prediction (EternaFold, RNAfold) as an independent view, assessing whether the Grassmannian subspace identified by RNA-FM aligns with physics-based structural ensembles. Second, full distillation-based factorization of RNA-FM to extract interpretable factors, following the pipeline established for GPT-2 [Tower, 2026c]. Third, application to additional RNA therapeutic targets (SARS-CoV-2 frameshifting element, DMD exon-skipping targets, SMA survival motor neuron) to test whether the bracket-norm and transportability metrics generalize across RNA drug targets.

6 Conclusion

We applied three geometric audits—bracket-norm correction, Grassmannian transportability, and weight factorization—to RNA-FM representations of the HTT CAG repeat region. The central finding is a categorical failure of naive transportability: RNA-FM embeddings undergo near-complete subspace rotation for any change in CAG repeat count, with no gradual degradation or clinically relevant threshold effect. Weight factorization confirms that the model’s internal structure admits low-rank decomposition (W_Q effective rank 86 in layer 11, W_Q – W_K overlap 0.81 in layer 3), providing a path toward mechanistic interpretability of RNA foundation models. These tools together establish a geometric validity check for RNA structural claims derived from foundation model embeddings, and quantify the transportability barrier that must be overcome for embedding-guided RNA-targeting drug design.

All code, data, and analysis results are available at <https://github.com/ElliottTower/causal-rna> under the MIT license.

References

- Tower, E. Bracket norm: Distinguishing intensive from extensive properties in high-dimensional biomedical feature spaces. *Zenodo*, 2026.
- Tower, E. Grassmannian cohort transfer: Quantifying subspace transportability across biomedical conditions. *Zenodo*, 2026.
- Tower, E. Factorization circuits: Reading transformer circuits from weight decompositions. *Zenodo*, 2026.
- Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., King, I., et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv:2204.00300*, 2022.
- de Mezer, M., Wojciechowska, M., Napierala, M., Sobczak, K., and Krzyzosiak, W. J. Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. *Nucleic Acids Research*, 39(9):3852–3863, 2011.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.
- MacDonald, M. E., Ambrose, C. M., Duyao, M. P., et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington’s disease chromosomes. *Cell*, 72(6):971–983, 1993.
- Rubinsztein, D. C., Leggo, J., Coles, R., et al. Phenotypic characterization of individuals with 30–40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36–39 repeats. *American Journal of Human Genetics*, 59(1):16, 1996.
- Sobczak, K., de Mezer, M., Michlewski, G., Krol, J., and Krzyzosiak, W. J. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Research*, 31(19):5469–5482, 2003.
- Tabrizi, S. J., Leavitt, B. R., Landwehrmeyer, G. B., et al. Targeting Huntingtin expression in patients with Huntington’s disease. *New England Journal of Medicine*, 380(24):2307–2316, 2019.
- Tabrizi, S. J., Estevez-Fraga, C., van Roon-Mom, W. M. C., et al. Potential disease-modifying therapies for Huntington’s disease: lessons learned and future opportunities. *The Lancet Neurology*, 21(7):645–658, 2022.
- Wolf, T., Debut, L., Sanh, V., et al. Transformers: State-of-the-art natural language processing. In *EMNLP: System Demonstrations*, 2020.