SAME ERROR, DIFFERENT GEOMETRY: RELATIONAL GENERALIZATION ON COMBINATORIAL LANDSCAPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard predictive error can be insufficient on structured combinatorial spaces because models with similar aggregate error can organize residuals differently over the underlying graph. This matters when predictions are used not only to estimate individual fitness values, but also to recover mutational effects, interactions, landscape structure, and search-relevant relationships. We formalize this distinction as *relational generalization*, separating pointwise predictive risk from first- and second-order graph-relative fidelity. Across four empirical RNA fitness landscapes and seven model classes, natural matched-risk comparisons showed limited first-order separation, motivating a controlled intervention that exactly matched held-out pointwise error while preserving learned residual organization. Despite identical held-out pointwise risk, learned residual directions retained substantial differences in first- and second-order relational error, with second-order error differing by up to $2.8\times$ within the same system and split. These differences altered reconstructed local-optimum and accessibility structure. Mutation-sign disagreements were concentrated among low- to moderate-margin effects, while high-margin effects were largely stable, consistent with a margin bound linking relational error to sign errors above a specified effect margin. Controlled spectral constructions on Domingo tRNA and *Tetrahymena thermophila* provided a complementary demonstration that equal aggregate error can coexist with different relational geometry and reconstructed topology. Relation-aware training further reduced relational error in both tested biological systems, although pointwise equivalence was not established. Sensitivity analysis showed that the main joint-preservation result was robust across a broad range of first- versus second-order weightings. Overall, our results show that aggregate predictive error alone does not determine relational geometry or structural fidelity, motivating relational generalization as a complementary criterion for evaluating and training models intended for structured biological inference and sequence search.

1 INTRODUCTION

Learned surrogates are commonly evaluated by pointwise predictive error, yet in structured combinatorial domains their predictions also induce local gradients, interaction patterns, peaks, accessible paths, and search trajectories (du Plessis et al., 2016; Jin, 2011). Fitness landscapes make this distinction explicit: genotypes are connected by mutational neighborhoods in which epistasis, ruggedness, and accessibility shape adaptation and search (Kauffman & Levin, 1987; Bank, 2022). Modern sequence models increasingly predict experimentally measured genotype-to-fitness maps from sparse observations (Shulgina et al., 2024; Sandhu et al., 2025; Hou et al., 2026) . A predictor may therefore achieve low error on individual fitness values while still misrepresenting the relations that determine landscape structure. This raises a basic question: *does aggregate predictive accuracy determine the relational and structural fidelity of the reconstructed landscape?*

This distinction matters whenever learned predictions are used for more than estimating individual values. In biological sequence modelling and other structured combinatorial problems, predictive models may be used to infer mutational effects and interactions, reconstruct landscape structure, compare candidate models, or guide search over candidate designs. If two models with similar predictive error recover these relations differently, conventional test error alone may favor a surrogate that is accurate pointwise but unreliable for the structure on which interpretation and search de-

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.

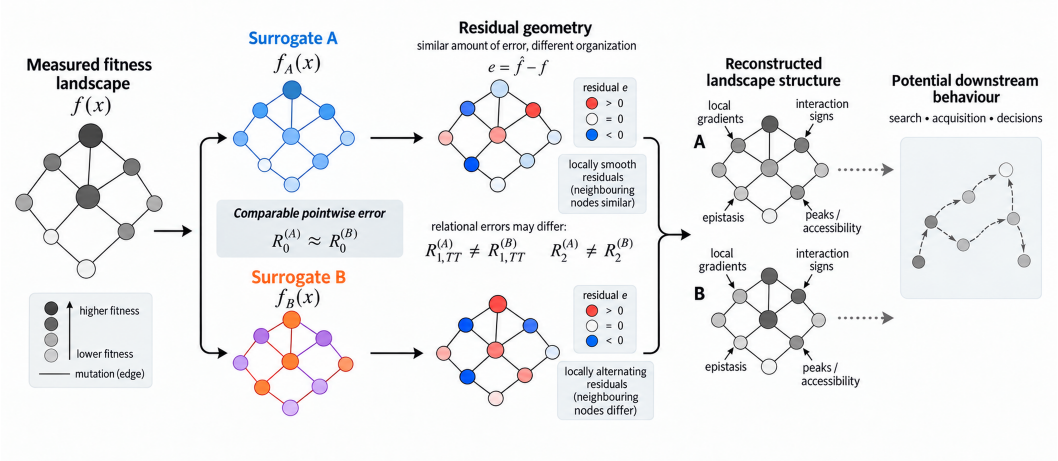


Figure 1: **Relational generalization.** Similar aggregate predictive error can have different graph-relative organization, producing different first- and second-order relational errors and different reconstructed landscape structure, with possible downstream implications.

pend. A complementary notion of relational fidelity may therefore improve how such models are evaluated, selected, and trained for structure-dependent scientific and optimization tasks.

We study the *residual geometry* of learned fitness models: the graph-relative organization of the residual field $e = \hat{f} - f$. Pointwise error constrains the magnitude of this field but not how it varies across neighboring genotypes or mutation squares. Two residual fields with the same norm can therefore induce different mutational effects, epistatic interactions, local optima, and accessible paths. We formalize this distinction as *relational generalization*, separating pointwise risk R_0 from first-order mutational error R_1 and second-order interaction error R_2. We further show that relational error has a direct structural interpretation: above a specified true-effect margin, R_1 and R_2 bound the number and, on fixed support, the fraction of mutation and interaction-sign errors, linking residual geometry to the local landscape structure. Relational generalization therefore provides a structural-fidelity diagnostic for surrogate models whose predictions are used to interpret or exploit combinatorial landscapes.

We make three main contributions. First, we formalize relational generalization and connect first- and second-order relational error to mutation- and interaction-sign recovery and local landscape structure. Second, we show across empirical and controlled matched-R_0 analyses that equal aggregate predictive error can coexist with different relational geometry and reconstructed landscape structure. In the empirical analysis, residual directions from 84 trained models are counterfactually rescaled to exactly equal held-out R_0, yet retain different held-out R_1 and R_2, different margin-dependent mutation-sign recovery, and different reconstructed local-optimum and accessibility structure. Controlled constructions on Domingo tRNA and *Tetrahymena thermophila* provide a complementary test by varying residual geometry while holding R_0 exactly fixed.

Third, we examine whether relational fidelity can be modified through learning. Relation-aware objectives alter first- and second-order fidelity in system-dependent ways and reduce an operational relational composite in both tested systems, although pointwise equivalence is not established. We also evaluate downstream sequential-decision relevance; the observed associations with relational quantities weaken substantially after adjustment and are therefore treated as exploratory. Overall, these results show that *pointwise accuracy alone is insufficient to characterize how prediction errors are organized over the landscape or how faithfully local landscape structure is reconstructed.*

2 Related Work

Fitness-landscape analysis and surrogate-assisted evolutionary computation study prediction and search over structured objective spaces (Kauffman & Levin, 1987; Pitzer & Affenzeller, 2012; Jin, 2005; 2011). Biological sequence models such as RNA-FM, RNAErnie, ERNIE-RNA, and Ri-

NALMo extend this setting to genotype-to-fitness prediction (Chen et al., 2022; Wang et al., 2024; Yin et al., 2025; Penić et al., 2025), while recent RNA and protein studies emphasize mutational effects, epistasis, and ruggedness (Shulgina et al., 2024; Sandhu et al., 2025; Hou et al., 2026). NABench broadens evaluation of nucleotide foundation models across many fitness-prediction assays (Li et al., 2025). Our question is complementary to representation ranking and benchmark coverage: we ask whether comparable aggregate predictive risk can conceal different relational and reconstructed landscape structure.

Beyond pointwise error, incomplete fitness maps can bias inferred epistasis and adaptive paths (Otwinowski & Plotkin, 2014), while minimum-epistasis interpolation and GraphFLA explicitly exploit or characterize landscape structure (Zhou & McCandlish, 2020; Huang et al., 2025). These approaches motivate evaluating properties of the reconstructed landscape rather than predictive error alone.

A related line of work studies interaction structure. Fourier–Walsh analyses decompose combinatorial functions by interaction order (Weinreich et al., 2018; Rahimimovassagh et al., 2026), while recent order-resolved frameworks such as ORBIT further distinguish interaction presence, representation accessibility, and functional recovery. Our graph-Fourier analysis addresses a different object: the spectral organization of the *prediction residual* on the mutational graph. It is therefore not interpreted as Walsh interaction order or as a substitute for order-resolved functional-recovery diagnostics. Instead, R_1, R_2, sign recovery, and residual-frequency organization provide complementary views of how prediction error is arranged over biologically meaningful relations.

Relation-aware training is closely connected to derivative-aware and graph-regularized learning. Sobolev training augments value supervision with derivative targets (Czarnecki et al., 2017), while Jacobian- and higher-derivative matching similarly constrain local functional behavior. Graph-Laplacian regularization and graph neural networks impose or learn neighborhood-dependent structure, and decision-focused learning optimizes models through downstream objectives (Zhou & Burges, 2008; Rahaman et al., 2019; Aghazadeh et al., 2021; Wilder et al., 2019). Our intervention is narrower: first- and second-order relation-preservation terms test whether relational fidelity can be modified while ordinary regression remains the primary predictive objective. The graph-smoothing control probes generic local smoothness but is not treated as equivalent to a dedicated graph-neural architecture or continuous Jacobian/Hessian supervision.

NABench and related benchmarks ask how models perform across biological prediction tasks, whereas ORBIT asks which interaction orders are functionally recovered. We study a complementary axis: whether the *geometry of residual error*, after controlling its aggregate magnitude, changes mutational, interaction, and reconstructed landscape structure. Direct comparison with Sobolev/Jacobian/Hessian-matching objectives, graph-neural baselines, and order-resolved diagnostics remains an important extension; the present experiments isolate the residual-geometry question using fixed model families, relation-aware objectives, and controlled matched-R_0 analyses.

3 Problem Formulation

A measured combinatorial landscape is $G = (V, E, f)$, where V contains genotypes, E admissible single-substitution edges, and $f(v)$ measured fitness. A predictor $\hat{f}$ induces the reconstructed landscape $\hat{G} = (V, E, \hat{f})$ and residual field $e = \hat{f} - f$. For a fixed landscape scale $s_G > 0$, evaluation nodes V_{eval}, edges E_{eval}, and mutation squares Q_{eval}, we define

$$R_0 = \frac{1}{s_G} \sqrt{\frac{1}{|V_{\text{eval}}|} \sum_{v \in V_{\text{eval}}} e(v)^2}, \tag{1}$$

$$R_1 = \frac{1}{s_G} \sqrt{\frac{1}{|E_{\text{eval}}|} \sum_{a \in E_{\text{eval}}} \left(\Delta_a \hat{f} - \Delta_a f\right)^2}, \tag{2}$$

$$R_2 = \frac{1}{s_G} \sqrt{\frac{1}{|Q_{\text{eval}}|} \sum_{q \in Q_{\text{eval}}} \left(\Delta_q \hat{f} - \Delta_q f\right)^2}. \tag{3}$$

Here Δ_a is the signed fitness difference across a single-substitution edge, while Δ_q is the four-corner second difference on a mutation square. Thus R_0 measures aggregate residual magnitude,

whereas R_1 and R_2 measure errors in first- and second-order relations. Our primary strictly held-out first-order endpoint, $R_{1,TT}$, uses edges whose two endpoints are held out; $R_{2,4\text{held}}$ analogously uses squares whose four vertices are held out. Evaluation supports are fixed independently of model predictions; complete support and normalization definitions are given in Appendix B.1.

Equal pointwise error does not determine relational error. On common first-order support, let B denote the oriented incidence matrix and $L = B^\top B$ the corresponding graph Laplacian. Then

$$\frac{R_1^2}{R_0^2} = \frac{|V_{\text{eval}}|}{|E_{\text{eval}}|} \frac{e^\top L e}{e^\top e}. \tag{4}$$

Thus fixing R_0 fixes the norm of the residual field but not its orientation relative to the graph. Residuals with equal pointwise magnitude can therefore have different R_1 when their spectral energy distributions induce different Laplacian Rayleigh quotients. The second-order analogue is governed by $e^\top C_Q^\top C_Q e$, where C_Q applies mutation-square second differences. We refer to this operator-relative organization of the residual field as its *residual geometry*. Because R_1 and graph-Laplacian Dirichlet energy are algebraically linked on common support, they are two descriptions of the same first-order structure rather than independent evidence. Spectral bounds, null-space cases, and higher-order extensions are given in Appendix A.1.

Relational error also constrains concrete structural mistakes. For $\gamma > 0$, define the set of first-order effects above a normalized margin

$$M_{1,\gamma} = \{a \in E_{\text{eval}} : |\Delta_a f| > \gamma s_G\},$$

and let $N_{1,\gamma}$ count edges in $M_{1,\gamma}$ for which the predicted and true mutation effects have different signs. Define $M_{2,\gamma}$ and $N_{2,\gamma}$ analogously for mutation-square second differences.

Proposition 1 (Margin-dependent sign control)**.** *For any predictor and any $\gamma > 0$,*

$$\frac{N_{1,\gamma}}{|E_{\text{eval}}|} \le \frac{R_1^2}{\gamma^2}, \qquad \frac{N_{2,\gamma}}{|Q_{\text{eval}}|} \le \frac{R_2^2}{\gamma^2}. \tag{5}$$

Consequently, among relations exceeding the margin,

$$\frac{N_{1,\gamma}}{|M_{1,\gamma}|} \le \frac{|E_{\text{eval}}|}{|M_{1,\gamma}|} \frac{R_1^2}{\gamma^2}, \qquad \frac{N_{2,\gamma}}{|M_{2,\gamma}|} \le \frac{|Q_{\text{eval}}|}{|M_{2,\gamma}|} \frac{R_2^2}{\gamma^2}. \tag{6}$$

Eq. (6) gives the corresponding conditional sign-error bound among relations that exceed the specified margin.

A sign reversal of a true relation with magnitude greater than γs_G requires relational error exceeding γs_G on that edge or square. Summing these required squared errors yields Eq. (5). The result is deterministic and requires no distributional assumption: relational error may concentrate around small-margin relations, but sufficiently large-margin mutation and interaction signs cannot be changed arbitrarily. The full proof is given in Appendix A.1.

Corollary for local landscape structure. Let H_γ denote the set of vertices whose incident mutation edges are all included in E_{eval} and whose true incident effects satisfy $|\Delta_a f| > \gamma s_G$. Let C_γ count vertices in H_γ whose local-optimum status differs between f and $\hat{f}$. A change in local-optimum status requires at least one incident sign error counted by $N_{1,\gamma}$, while each erroneous edge is incident to at most two vertices. Hence

$$C_\gamma \le 2N_{1,\gamma}, \qquad \frac{C_\gamma}{|H_\gamma|} \le \frac{2|E_{\text{eval}}|}{|H_\gamma|} \frac{R_1^2}{\gamma^2}. \tag{7}$$

Thus first-order relational error directly constrains changes in fully supported high-margin local-optimum status. This result does not imply an analogous universal bound for global descriptors such as accessibility, which can depend on many coupled local decisions; such reconstructed properties are therefore evaluated explicitly.

Together, Eqs. (1)–(4), Proposition 1, and Eq. (7) distinguish residual magnitude, relational sign errors, and changes in reconstructed local landscape structure. Figure 1 summarizes this distinction; the empirical analyses below test it using learned residual fields, controlled residual directions, and relation-aware training.

4 METHODS AND EXPERIMENTS

Data, splits, and common evaluation. We use four measured RNA fitness landscapes—CPEB3, Domingo tRNA, F1*U, and Tetrahymena—containing 21,322, 4,176, 27,903, and 64,020 genotypes, respectively, with the four-site GB1 protein landscape providing supporting cross-domain evidence. Train, validation, and test partitions are fixed before relational evaluation, and model selection uses only train/validation data. Unless stated otherwise, pointwise and relational errors use the training-fitness scale

$$s_G = \mathrm{SD}\{f(v) : v \in V_{\mathrm{train}}\}.$$

The primary strictly held-out relational endpoints are $R_{1,TT}$, using single-substitution edges whose two endpoints are test nodes, and $R_{2,4\mathrm{held}}$, using mutation squares whose four vertices are test nodes. Evaluation supports are determined from the observed landscape and data split independently of model predictions. Support, normalization, and provenance details are given in Appendix B.

Learned-model benchmark and residual-structure analysis. The main benchmark contains seven model families: One-hot Ridge, Pairwise Ridge, a Scratch Transformer, RNA-FM, RNAErnie, ERNIE-RNA, and RiNALMo, each evaluated on four RNA systems with three seeds, giving 84 fits. The predefined natural matched-risk comparison pairs One-hot Ridge and Scratch within system and seed when $|\Delta R_0| \leq 0.05$, with $R_{1,TT}$ as its primary endpoint.

A learned-structure analysis uses One-hot Ridge, Pairwise Ridge, and Scratch on Domingo with five seeds and two fixed holdout modes (30 runs), evaluated on common mutation-square support using R_2, second-difference sign recovery, and sign-epistasis diagnostics. A graph-Fourier analysis decomposes each full-landscape residual field over the Domingo graph Laplacian, separating the constant component from non-constant energy in four fixed $\lambda_{\max}$-relative frequency bands. These bands describe graph-Fourier residual-frequency organization and are not interpreted as Walsh–Fourier interaction order. Paired analyses and spectral details are given in Appendix C.3.

Learned-direction matched-R_0 calibration. To separate learned residual organization from magnitude, we counterfactually rescale each of the 84 fitted-model residual fields without retraining. For model m, let:

$$e_m(v) = \hat{f}_m(v) - f(v)$$

denote its complete residual field and let $\mathrm{RMSE}_{\mathrm{test}}(e_m)$ denote its residual RMSE on held-out nodes. For target normalized point-wise error $\tau \in \{0.10, 0.20, 0.30\}$, we define:

$$c_{m,\tau} = \frac{\tau s_G}{\mathrm{RMSE}_{\mathrm{test}}(e_m)}, \qquad \tilde{f}_{m,\tau} = f + c_{m,\tau} e_m. \tag{8}$$

In Eq. (8), $c_{m,\tau} > 0$, so the transformation preserves the direction and component-wise sign pattern of the learned residual field while forcing test-node R_0 to equal τ exactly. Held-out labels are used only to impose this equality constraint; they are not used to train, select, or tune any predictor. These are counterfactual reconstructions from magnitude-normalized learned residual directions, not separately trained models that naturally attained equal R_0.

For every calibrated reconstruction we evaluate $R_{1,TT}$ on held-out test–test edges and $R_{2,4\mathrm{held}}$ on mutation squares whose four vertices are held out. We additionally evaluate first-order mutation-sign accuracy on test–test edges at normalized true-effect margins $\gamma \in \{0.10, 0.25, 0.50, 1.00\}$, defined from the true landscape and therefore prediction-independent. The 84 residual directions at three common amplitudes yield 252 calibrated reconstructions. This secondary analysis is descriptive; seeds quantify within-system stability and amplitudes are conditions rather than independent replicates.

Structural reconstruction and controlled residual geometry. For each magnitude-normalized learned residual field, we reconstruct the full fitness landscape and evaluate GraphFLA structural descriptors on the observed single-substitution graph. Local-optimum ratio is the primary topological endpoint, with global-optimum accessibility evaluated where defined. These are counterfactual full-landscape reconstruction diagnostics, not held-out topology-generalization endpoints; held-out generalization is assessed through $R_{1,TT}$, $R_{2,4\mathrm{held}}$, and mutation-sign recovery. Graph construction, neutral-edge handling, accessibility definitions, and the F1*U isolated-node case are documented in Appendix C.10.

We separately construct deterministic matched-R_0 residual fields on Domingo and Tetrahymena to isolate the mechanism under direct control of residual geometry. On Domingo we use low- and high-frequency Laplacian directions; on Tetrahymena we use corresponding fixed smooth/low-frequency and alternating/high-frequency directions. Each direction is normalized to $R_0 \in \{0.10, 0.20, 0.30\}$ exactly, after which R_1, R_2, and the same structural descriptors are evaluated. These constructions test the existence and portability of the mechanism on empirical landscapes, not its biological prevalence; amplitudes are conditions rather than independent replicates. Because R_1 and graph-Laplacian Rayleigh energy are algebraically linked on common support, they are not treated as independent evidence. Details are in Appendix C.6.

Relation-aware training and weighting sensitivity. Relation-aware training is evaluated on CPEB3 and Tetrahymena using a compact RNA encoder pretrained from scratch on a versioned RNACentral-derived corpus and frozen before downstream fitting. Three data splits and two model seeds per split yield 192 fits across conventional regression, joint first- and second-order relation-preservation, graph-smoothing, contrastive, and structured-control families. Candidate selection uses validation data only, candidates must satisfy:

$$R_0^{\mathrm{val}} \leq R_0^{\mathrm{val}}(\mathrm{regression}) + 0.05,$$

after which the admissible candidate within each objective family minimizing the predefined operational composite

$$R_{\mathrm{rel}}^{\mathrm{val}} = R_1^{\mathrm{val}} + R_2^{\mathrm{val}}$$

is selected. All relation-preservation targets used for fitting are constructed exclusively from training-labeled nodes and relations; test–test edges and four-held-out mutation squares are reserved for evaluation and never used in fitting or model selection. Equal weighting is the predefined operational selection convention, not a claim that first- and second-order errors are biologically commensurate. Pointwise equivalence is evaluated with the predefined ± 0.05 TOST margin at the biological-system level.

A subsequent objective decomposition adds 36 second-order-only fits while leaving the original 192-fit campaign unchanged. A Pareto audit compares already selected validation-eligible objectives in $(\Delta R_0, \Delta R_{\mathrm{rel}})$ space without reselection. We also recompute the fixed held-out contrasts under:

$$R_{\mathrm{rel}}(\alpha) = \alpha R_1 + (1-\alpha) R_2, \qquad \alpha \in [0,1], \tag{9}$$

Eq. (9) assesses sensitivity to the equal-weight convention using existing selected predictions only; no model is retrained or reselected, and no new confirmatory inference is introduced. Architecture, eligibility, decomposition, Pareto, and weighting details are given in Appendices B.5, B.6, and C.8.

Downstream evaluation and inferential status. Downstream relevance is evaluated on CPEB3 and Tetrahymena using 120 matched sequential-acquisition trajectories under a fixed global-greedy primary policy. The predefined intervention criterion is a system-level decision improvement of at least 0.05. A separate association analysis adds $R_{1,TT}$ and $R_{2,4\mathrm{held}}$ to conventional predictive summaries, with adjustment for biological system and model role as a sensitivity analysis. Biological systems are the independent units for cross-system inference. Repeated splits and seeds quantify within-system stability, while calibrated amplitudes are conditions rather than replicates. Statistical procedures and downstream sensitivities are given in Appendices B.2 and C.9.

5 Results and Discussion

5.1 Natural matching provides a negative boundary; magnitude-normalized learned errors reveal distinct relational geometry

The predefined natural matched-risk comparison did not establish the hypothesized large first-order separation. Only three Tetrahymena One-hot–Scratch pairs satisfied $|\Delta R_0| \leq 0.05$, with mean $|\Delta R_{1,TT}| = 0.00481$ and maximum 0.00592, far below the predefined 0.10 separation criterion. Broader matching across the model roster, conducted after the analysis freeze, produced the same qualitative boundary (Appendix C.2). Thus the natural comparisons constrain the size of spontaneously occurring matched-risk separation, but do not establish that pointwise error determines relational geometry.

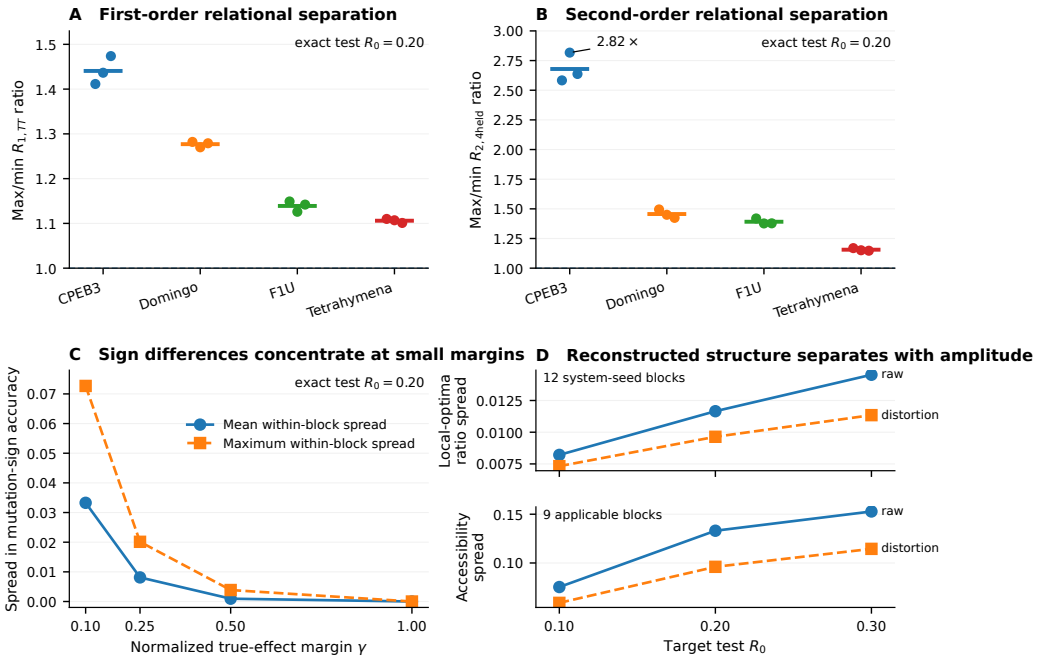


Figure 2: **Learned-direction matched-R_0 analysis.** Frozen residual directions are rescaled to identical held-out pointwise error. **A,** Within-block max/min ratios of held-out $R_{1,TT}$ at exact test $R_0 = 0.20$. **B,** Corresponding max/min ratios of $R_{2,4\text{held}}$. **C,** Within-block spread in first-order mutation-sign accuracy across normalized true-effect margins γ; differences concentrate at lower margins. **D,** Spread in reconstructed local-optimum ratio and global-optimum accessibility across target R_0, shown as raw spread and spread in absolute distortion from the true landscape. Relational quantities are evaluated on held-out support, whereas topology is a full-landscape reconstruction diagnostic.

To isolate residual organization from magnitude directly, we performed a learned-direction matched-R_0 analysis over the 84 fitted models spanning four RNA landscapes, seven model families, and three seeds. For each fit, the learned residual field was multiplied by a single positive scalar so that its held-out pointwise error equalled exactly $R_0 \in \{0.10, 0.20, 0.30\}$. This transformation preserves the learned residual direction while removing differences in its magnitude; it is therefore a counterfactual calibration of learned residual fields, not a comparison of independently trained models that naturally attained identical R_0. The analysis is reported without confirmatory inference; seeds quantify within-system stability, and calibrated amplitudes are experimental conditions rather than independent replicates.

Relational differences survive exact pointwise normalization. At $R_0 = 0.20$, the max/min $R_{1,TT}$ ratio across learned residual directions within the same system and seed ranges from approximately $1.10\times$ to $1.47\times$. Second-order separation is larger: the corresponding $R_{2,4\text{held}}$ ratio reaches $2.82\times$ on CPEB3, $1.50\times$ on Domingo, $1.42\times$ on F1*U, and $1.17\times$ on Tetrahymena. Held-out square support is comparatively sparse for CPEB3, Domingo, and F1*U and dense for Tetrahymena; complete support counts and system–seed results are reported in Appendix C.10.

The structural consequences are margin dependent. At $R_0 = 0.20$ and normalized first-order effect margin $\gamma = 0.10$, the mean within-block spread in mutation-sign accuracy is 0.0333 (maximum 0.0727). The mean spread falls to 0.0081 at $\gamma = 0.25$, below 0.001 at $\gamma = 0.50$, and vanishes at $\gamma = 1$. This pattern is consistent with Proposition 1, which bounds sign errors above a specified true-effect margin in terms of relational error: small-margin mutation effects are more vulnerable to residual organization, whereas large-margin signs are highly stable.

The same magnitude-normalized learned directions also reconstruct different landscape structure. At $R_0 = 0.20$, the mean within-system–seed spread across model families is 0.0117 in local-optimum

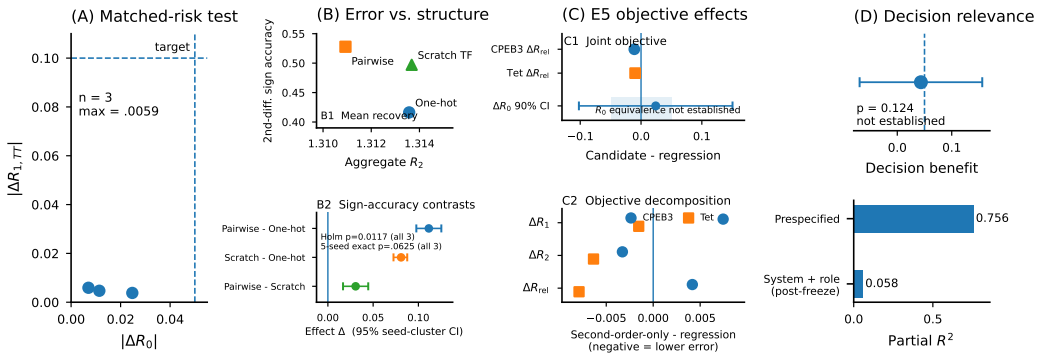


Figure 3: **Complementary empirical analyses. A,** Natural matched-risk comparisons provide a negative boundary on large first-order separation. **B,** Similar aggregate R_2 can mask different interaction-sign recovery and residual organization. **C,** Relation-aware training changes first- and second-order relational fidelity, although pointwise equivalence is not established. **D,** Downstream relational association attenuates strongly after controlling for biological system and model role.

ratio and 0.133 in global-optimum accessibility across the nine blocks where accessibility is defined. Relative to the true landscapes, the corresponding mean spreads in absolute distortion are 0.0096 and 0.096. These structural spreads increase descriptively as the calibrated residual amplitude increases from $R_0 = 0.10$ to 0.30. Exact equality of held-out pointwise error therefore does not determine either held-out relational fidelity or the structural fidelity of the reconstructed landscape.

5.2 Complementary learned and controlled evidence supports the mechanism

Similar aggregate second-order error can obscure differences in learned structure. On Domingo, One-hot Ridge, Pairwise Ridge, and Scratch have mean held-out $R_2 = 1.313574$, 1.310930, and 1.313679, respectively, while their second-difference sign accuracies are 0.416, 0.528, and 0.497. One-hot and Scratch are particularly informative: their mean R_2 differs by only 1.05×10^{-4}, yet their sign accuracy differs by 0.0809. Their full-landscape residual organization also differs sharply: One-hot and Scratch place 0.262 and 0.515 of nonconstant residual energy in the lowest fixed graph-Fourier band, with the direction of this contrast preserved across all seed–holdout configurations and seed averages. Nearly identical aggregate second-order error can therefore coexist with markedly different interaction recovery and residual-frequency organization. These repeated configurations quantify within-landscape stability rather than independent biological replication; complete interaction-sign, sign-epistasis, spectral, and reconstruction results are reported in Appendix C.3.

Controlled matched-R_0 constructions isolate residual geometry. A complementary construction varies residual direction while fixing its pointwise norm exactly. On Domingo, smooth and high-frequency residual directions have Rayleigh quotients 1.59245 and 23.79003, a $14.94\times$ separation. At $R_0 = 0.20$, they yield $R_1 = 0.1005$ versus 0.3884 and $R_2 = 0.0188$ versus 0.7687, together with a 0.0278 difference in reconstructed local-optimum ratio. Because R_1 and the graph-Laplacian Rayleigh quotient are algebraically linked, they describe the same first-order mechanism rather than independent evidence.

The same mechanism is observed on Tetrahymena. At exactly $R_0 = 0.20$, smooth and alternating/high-frequency residuals yield $R_1 = 0.1140$ versus 0.3301 and $R_2 = 0.0105$ versus 0.5491, while reconstructed local-optimum ratios are 0.0160 and 0.0226. Accessibility also changes under both constructions, although closeness to the true landscape is descriptor dependent. These deterministic analyses provide a complementary demonstration of the mechanism, not an estimate of biological prevalence or independent replication. The full amplitude sweeps and mechanism figure are reported in Appendix C.6.

5.3 Relation-aware training changes relational fidelity, with order-specific trade-offs

Joint first- and second-order preservation reduced the predefined operational composite $R_{\mathrm{rel}} = R_1 + R_2$ relative to regression in both tested systems: $\Delta R_{\mathrm{rel}} = -0.0110$ on CPEB3 and -0.0097 on Tetrahymena. The mean system-level point-wise difference was $\Delta R_0 = +0.0241$, but the hierarchical TOST 90% confidence interval $[-0.1022, 0.1505]$ did not establish point-wise equivalence under the prespecified ± 0.05 margin. Failure to establish equivalence does not, by itself, establish point-wise harm.

The two relational orders capture distinct aspects of model error. Second-order-only preservation reduces R_2 in both systems (-0.0033 on CPEB3 and -0.0064 on Tetrahymena), while its effect on R_1 changes sign ($+0.0074$ and -0.0015, respectively). We therefore treat R_{rel} as an operational equal-weight summary rather than a canonical measure of relational structure and report its components separately. A reporting sensitivity analysis using:

$$R_{\mathrm{rel}}(\alpha) = \alpha R_1 + (1 - \alpha) R_2$$

shows that the joint-versus-regression contrast remains negative across $\alpha \in [0, 1]$ for the common-eligible CPEB3 comparison and remains negative on Tetrahymena until essentially all weight is placed on R_1 ($\alpha \approx 0.999$). Other preservation-arm contrasts are more weight-sensitive, reinforcing the need to retain R_1 and R_2 separately. This sensitivity involves no retraining, reselection, or new confirmatory inference. The graph-smoothing control, validation eligibility, objective decomposition, weighting sensitivity, and descriptive Pareto audit are reported in Appendices C.7 and C.8.

5.4 Downstream evidence remains exploratory

Under the predefined global-greedy policy, the mean system-level decision benefit of the relational intervention was 0.0434, below the required 0.05 threshold ($p = 0.124$), so the decision-improvement criterion was not established. In a separate association analysis, adding $R_{1,TT}$ and $R_{2,4\mathrm{held}}$ increased in-sample explanatory power, but the corresponding partial R^2 fell from 0.756 to 0.058 after controlling for biological system and model role. We therefore treat this as exploratory evidence of decision-relevant association within the evaluated setting, not as evidence of a large independent or causal downstream effect (Appendix C.9).

The learned-direction analysis is a counterfactual calibration performed after model fitting; it does not imply that independently trained models naturally attain identical R_0. Topology is evaluated as full-landscape reconstruction, calibrated amplitudes are conditions rather than replicates, and repeated seeds quantify within-system stability. Within these limits, equal aggregate predictive error does not determine relational or reconstructed structural fidelity.

6 Conclusion

Relational generalization evaluates learned surrogates not only by the magnitude of their prediction error, but also by how that error is organized over a combinatorial landscape. Natural matched-risk comparisons place a negative boundary on large naturally occurring first-order separation, whereas counterfactual normalization of residual directions from trained models shows that exactly equal held-out R_0 can coexist with different R_1, R_2, mutation-sign recovery, and reconstructed landscape structure. Margin-based bounds further connect relational error to high-margin sign errors and local-optimum instability, while controlled constructions isolate residual geometry as a mechanism underlying these differences.

Relation-aware training further shows that relational fidelity can be modified, although first- and second-order effects exhibit system-dependent trade-offs and pointwise equivalence was not established. Downstream evidence remains exploratory. Future work should identify the learning mechanisms that produce unfavorable residual geometry and develop training or acquisition strategies that improve structural fidelity while preserving pointwise accuracy.

Overall, aggregate predictive error alone does not determine relational geometry or reconstructed structural fidelity. Pointwise accuracy remains essential, but it is not a complete description of a learned landscape.

AI USE STATEMENT

Generative AI-assisted tools were used in the preparation and review of this manuscript. Language models were used solely for language editing and manuscript presentation, including grammar, spelling, clarity, readability, LaTeX formatting, and consistency of cross-references. Google Gemini was also used to generate a conceptual figure based on an author-provided description and prompts; the resulting image was reviewed by the authors before inclusion.

In addition, the Stanford Agentic Reviewer LLM (paperreview.ai/review) and the OpenReview.net LLM were used at stages of manuscript preparation to obtain automated feedback on the paper. Some of this feedback informed subsequent revisions, but all suggested changes were independently assessed, selected, and implemented by the authors.

Generative AI tools were not used to develop the research question or hypotheses, design the methodology or experiments, generate or analyze data, independently interpret results, or formulate scientific conclusions. All scientific content, analyses, results, interpretations, and conclusions were produced and verified by the authors. All AI-assisted outputs and edits were reviewed and approved by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work is a computational study based on previously published and publicly accessible biological sequence to function datasets. No new experiments involving human participants, animals, clinical samples, or personally identifiable information were conducted or collected as part of this study. The work focuses on evaluating the relational and structural fidelity of predictive models on combinatorial fitness landscapes rather than on designing or experimentally validating biological sequences. Although methods for fitness prediction and surrogate-based search may have broader applications in biological design, the present study is methodological and does not target pathogenicity, virulence, or other harmful biological properties. All datasets and external models are cited to their original sources, and the analyses are reported with their stated statistical and experimental limitations.

REPRODUCIBILITY STATEMENT

Pretraining data for the relation-aware-training RNA encoder were assembled from RNACentral.[1] Downstream evaluation used experimentally measured landscapes from the CPEB3 ribozyme,[2] Domingo yeast tRNA,[3] F1*U ligase ribozyme,[4] and *Tetrahymena thermophila* group-I intron;[5] the GB1 protein landscape[6] provided supporting cross-domain evidence.

The anonymous repository (`https://anonymous.4open.science/r/SAME-ERROR-DIFFERENT-GEOMETRY-RELATIONAL-GENERALIZATION-ON-COMBINATORIAL-LANDSCAPES-7357/README.md`) provides the code, experiment configurations, frozen split definitions, exact edge and mutation-square supports, normalization and evaluation scripts, integrity checks, machine-readable manifests, artifact hashes, and model/corpus provenance required to reconstruct the reported analyses. It also includes the fixed relation-aware-training checkpoint and a lightweight API for computing R_1 and R_2 from user-supplied predictions together with arbitrary edge and mutation-square supports. The independently pretrained RNA corpus and final checkpoint are versioned and SHA-256 hashed, with audit records documenting validation-overlap removal, downstream-benchmark containment checks, and the sampled near-homology audit. External pretrained encoders are evaluated frozen under the common downstream protocol. Further preprocessing, split, architecture, computational-environment, metric implementation, and inferential details are provided in (Appendix **??**) , with additional sensitivity analyses in (Appendix C).

[1] `https://rnacentral.org/`
[2] `https://gitlab.com/bsu/biocompute-public/ml-ribo-predict.git`
[3] `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99418`
[4] `https://doi.org/10.5281/zenodo.6945203`
[5] `https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA636762`
[6] `https://doi.org/10.7554/eLife.16965.024`

Author Contributions

Author 1 and Author 2 contributed equally to this work and collaborated throughout the entire study, including the development of the research question, study design, methodology, implementation, analysis, interpretation of results, and preparation of the manuscript.

Acknowledgments

Acknowledgments have been omitted from the anonymous submission to preserve author anonymity and will be included in the final version if appropriate.

References

Amirali Aghazadeh, Hunter Nisonoff, Orhan Ocal, David H. Brookes, Yijie Huang, O. Ozan Koyluoglu, Jennifer Listgarten, and Kannan Ramchandran. Epistatic net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. *Nature Communications*, 12:5225, 2021. doi: 10.1038/s41467-021-25371-3.

Claudia Bank. Epistasis and adaptation on fitness landscapes. *Annual Review of Ecology, Evolution, and Systematics*, 53:457–479, 2022. doi: 10.1146/annurev-ecolsys-102320-112153.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Louis du Plessis, Gabriel E. Leventhal, and Sebastian Bonhoeffer. How good are statistical models at approximating complex fitness landscapes? *Molecular Biology and Evolution*, 33(9):2454–2468, 2016. doi: 10.1093/molbev/msw097.

Chao Hou, Di Liu, Aziz Zafar, and Yufeng Shen. Understanding language model scaling for protein fitness prediction. *Nature Computational Science*, 6(7):778–788, 2026. doi: 10.1038/s43588-026-01010-z.

Mingyu Huang, Shasha Zhou, and Ke Li. Augmenting biological fitness prediction benchmarks with landscape features from GraphFLA. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yaochu Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing*, 9(1):3–12, 2005. doi: 10.1007/s00500-003-0328-5.

Yaochu Jin. Surrogate-assisted evolutionary computation: Recent advances and future challenges. *Swarm and Evolutionary Computation*, 1(2):61–70, 2011. doi: 10.1016/j.swevo.2011.05.001.

Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45, 1987. doi: 10.1016/S0022-5193(87)80029-2.

Zhongmin Li, Runze Ma, Jiahao Tan, Chengzi Tan, and Shuangjia Zheng. NABench: Large-scale benchmarks of nucleotide foundation models for fitness prediction. *arXiv preprint arXiv:2511.02888*, 2025.

Jakub Otwinowski and Joshua B. Plotkin. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences*, 111(22):E2301–E2309, 2014. doi: 10.1073/pnas.1400849111.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, et al. RiNALMo: general-purpose RNA language models can generalize well on structure prediction tasks. *Nature Communications*, 16:5671, 2025. doi: 10.1038/s41467-025-60872-5.

Erik Pitzer and Michael Affenzeller. A comprehensive survey on fitness landscape analysis. In János Fodor, Ryszard Klempous, and Carmen Paz Suarez Araujo (eds.), *Recent Advances in Intelligent Engineering Systems*, volume 378 of *Studies in Computational Intelligence*, pp. 161–191. Springer, 2012. doi: 10.1007/978-3-642-23229-9_8.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5301–5310, 2019.

Maryam Rahimimovassagh, Ivan Garibay, and Niloofar Yousefi. Beyond tokens: Probing higher-order epistasis in learned protein representations. *arXiv preprint arXiv:2608.24953*, 2026.

Mahakaran Sandhu, Adam C. Mater, Dana S. Matthews, Matthew A. Spence, Artem A. Lenskiy, and Colin Jackson. Investigating the determinants of performance in machine learning for protein fitness prediction. *Protein Science*, 34(8):e70235, 2025. doi: 10.1002/pro.70235.

Yekaterina Shulgina, Marena I. Trinidad, Conner J. Langeberg, Hunter Nisonoff, Seyone Chithrananda, Petr Skopintsev, et al. RNA language models predict mutations that improve RNA function. *Nature Communications*, 15:10627, 2024. doi: 10.1038/s41467-024-54812-y.

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*, 6:548–557, 2024. doi: 10.1038/s42256-024-00836-4.

Daniel M. Weinreich, Yinghong Lan, Jacob Jaffe, and Robert B. Heckendorn. The influence of higher-order epistasis on biological fitness landscape topography. *Journal of Statistical Physics*, 172(1):208–225, 2018. doi: 10.1007/s10955-018-1975-3.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

Weijie Yin, Zhaoyu Zhang, Shuo Zhang, et al. ERNIE-RNA: an RNA language model with structure-enhanced representations. *Nature Communications*, 16:10076, 2025. doi: 10.1038/s41467-025-64972-0.

Dengyong Zhou and Christopher J. C. Burges. High-order regularization on graphs. In *Proceedings of the 6th International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

Juannan Zhou and David M. McCandlish. Minimum epistasis interpolation for sequence-function relationships. *Nature Communications*, 11:1782, 2020. doi: 10.1038/s41467-020-15512-5.

APPENDIX

A THEORY AND FORMAL DERIVATIONS

A.1 MATCHED POINTWISE RISK AND RELATIONAL AMPLIFICATION

Consider a common evaluation node set V containing n nodes and residual vector $e = \hat{f} - f$. Let E contain m evaluated mutation edges. From the definitions of normalized pointwise and first-order relational error,

$$R_0^2 = \frac{e^\top e}{s_G^2 n}, \qquad R_1^2 = \frac{e^\top L e}{s_G^2 m},$$

where $L = B^\top B$ is the Laplacian associated with the evaluated edge incidence matrix B. Therefore,

$$\frac{R_1^2}{R_0^2} = \frac{n}{m} \frac{e^\top L e}{e^\top e}.$$

Defining the first-order residual Rayleigh quotient

$$\omega_1(e) = \frac{e^\top L e}{e^\top e},$$

gives

$$\frac{R_1^2}{R_0^2} = \frac{n}{m}\omega_1(e).$$

Consequently, for two predictors evaluated on the same support with $R_0^{(A)} = R_0^{(B)} > 0$,

$$\frac{R_1^{(A)}}{R_1^{(B)}} = \sqrt{\frac{\omega_1(e_A)}{\omega_1(e_B)}},$$

whenever the denominator is nonzero. Matching pointwise error therefore fixes residual magnitude but not first-order relational error; the latter is determined by the orientation of residual energy relative to the graph Laplacian.

Let $\lambda_{\max}(L)$ denote the largest Laplacian eigenvalue. Then

$$0 \le \frac{R_1}{R_0} \le \sqrt{\frac{n}{m}\lambda_{\max}(L)}.$$

If the residual is restricted to the orthogonal complement of $\ker(L)$, the smallest positive eigenvalue provides the corresponding lower bound. In particular, componentwise constant residuals lie in $\ker(L)$: they may contribute to pointwise error while producing zero first-order relational error.

Second-order relations. Let Q denote a collection of q mutation squares, and define the square-difference operator C_Q such that, for $q = (x, x^{(a)}, x^{(b)}, x^{(ab)})$,

$$(C_Q e)_q = e(x^{(ab)}) - e(x^{(a)}) - e(x^{(b)}) + e(x).$$

Then

$$R_2^2 = \frac{e^\top C_Q^\top C_Q e}{s_G^2 q}.$$

Writing

$$H_Q = C_Q^\top C_Q, \qquad \omega_2(e) = \frac{e^\top H_Q e}{e^\top e},$$

gives

$$\frac{R_2^2}{R_0^2} = \frac{n}{q}\omega_2(e).$$

Thus equal pointwise residual energy also fails to determine second-order relational error unless the corresponding second-order Rayleigh quotients are equal. The null space of C_Q consists of residual patterns with zero second differences on the evaluated squares; on complete product supports, this includes constant and purely additive components.

Margin-dependent sign bound and local-optimum corollary. Let E_{eval} denote the evaluated first-order edge set and define

$$M_{1,\gamma} = \{a \in E_{\text{eval}} : |\Delta_a f| > \gamma s_G\}.$$

Let $N_{1,\gamma}$ be the number of edges in $M_{1,\gamma}$ on which the predicted and true mutational effects have opposite signs. For edge a, set

$$\delta_a = \Delta_a \hat{f} - \Delta_a f = \Delta_a e.$$

If the sign is reversed on an edge in $M_{1,\gamma}$, then the prediction must cross zero relative to the true effect, and therefore

$$|\delta_a| \ge |\Delta_a f| > \gamma s_G.$$

Consequently,

$$N_{1,\gamma}\gamma^2 s_G^2 \le \sum_{\substack{a \in M_{1,\gamma} \\ \text{sign error}}} \delta_a^2 \le \sum_{a \in E_{\text{eval}}} \delta_a^2 = |E_{\text{eval}}| s_G^2 R_1^2,$$

and hence

$$\frac{N_{1,\gamma}}{|E_{\text{eval}}|} \le \frac{R_1^2}{\gamma^2}.$$

Whenever $|M_{1,\gamma}| > 0$, the corresponding conditional sign-error fraction satisfies

$$\frac{N_{1,\gamma}}{|M_{1,\gamma}|} \leq \frac{|E_{\text{eval}}|}{|M_{1,\gamma}|} \frac{R_1^2}{\gamma^2}.$$

The same argument applies to second differences. If Q_{eval} is the evaluated square set,

$$M_{2,\gamma} = \left\{ q \in Q_{\text{eval}} : |\Delta_q^2 f| > \gamma s_G \right\},$$

and $N_{2,\gamma}$ counts second-difference sign reversals on this set, then

$$\frac{N_{2,\gamma}}{|Q_{\text{eval}}|} \leq \frac{R_2^2}{\gamma^2}, \qquad \frac{N_{2,\gamma}}{|M_{2,\gamma}|} \leq \frac{|Q_{\text{eval}}|}{|M_{2,\gamma}|} \frac{R_2^2}{\gamma^2},$$

whenever the conditional denominator is nonzero.

For the local-optimum consequence, let H_γ contain vertices whose entire incident mutation neighborhood is evaluated and whose true incident mutation effects all have magnitude greater than γs_G. Let C_γ be the number of vertices in H_γ whose strict local-optimum status changes after reconstruction. Any such change requires at least one incident first-order sign error. Because a mutation edge is incident to at most two vertices,

$$C_\gamma \leq 2N_{1,\gamma},$$

and therefore, whenever $|H_\gamma| > 0$,

$$\frac{C_\gamma}{|H_\gamma|} \leq \frac{2|E_{\text{eval}}|}{|H_\gamma|} \frac{R_1^2}{\gamma^2}.$$

If a relevant margin set or fully supported vertex set is empty, the corresponding conditional statement is vacuous. These bounds control margin-separated local sign errors and the resulting strict local-optimum changes; no analogous universal bound is claimed here for global accessibility, which is evaluated empirically.

Isotropic residual baseline. If residual direction is isotropic conditional on fixed $\|e\|_2$, then for any symmetric matrix A,

$$\mathbb{E}\left[\frac{e^\top A e}{e^\top e}\right] = \frac{\operatorname{tr}(A)}{n}.$$

Because $\operatorname{tr}(L) = 2m$,

$$\mathbb{E}\left[\frac{R_1^2}{R_0^2}\right] = 2.$$

Each row of C_Q contains four coefficients of magnitude one, so $\operatorname{tr}(C_Q^\top C_Q) = 4q$, yielding

$$\mathbb{E}\left[\frac{R_2^2}{R_0^2}\right] = 4.$$

These values provide a reference corresponding to isotropically oriented residual energy rather than a biological performance threshold.

Higher-order extension. More generally, let D_k compute kth-order discrete differences on a declared collection S_k of valid k-dimensional combinatorial faces. Then

$$R_k^2 = \frac{e^\top D_k^\top D_k e}{s_G^2 |S_k|}, \qquad \frac{R_k^2}{R_0^2} = \frac{n}{|S_k|} \frac{e^\top D_k^\top D_k e}{e^\top e}.$$

For standard kth-order finite differences, each row contains 2^k coefficients of magnitude one. Under the same isotropic residual-direction null,

$$\mathbb{E}\left[\frac{R_k^2}{R_0^2}\right] = 2^k.$$

The R_1 and R_2 quantities used in this work are therefore the first two members of a more general hierarchy of graph-relative residual operators.

Table 1: Supplementary overview of experimental components and inferential roles.

Study	System(s)	Scale	Purpose	Status
E1	4 RNA landscapes	84 runs	Primary relational benchmark	Frozen primary
StructRFM	4 RNA landscapes	12 runs	Model-roster sensitivity	Post-freeze
E3	Domingo tRNA	30 runs	Higher-order and residual-spectrum comparison	Supporting; spectral analysis post-freeze
Controlled matched-R_0	Domingo tRNA, Tetrahymena	12 constructions	Deterministic exact-error geometry/topology mechanism	Post-freeze controlled
Learned-direction matched-R_0	4 RNA landscapes	84 residual directions $\times$ 3 amplitudes $=$ 252 reconstructions	Exact pointwise-magnitude normalization of learned residual geometry	Post-freeze descriptive counterfactual
E5	CPEB3, Tetrahymena	192 fits	Relational intervention	Frozen confirmatory
E5 decomposition/Pareto	CPEB3, Tetrahymena	36 new fits + fixed selections	First/second-order mechanism, weighting sensitivity, and trade-off structure	Post-freeze ablation/audit
GraphFLA	4 RNA landscapes	4 true + 84 E1 profiles	Baseline landscape reconstruction audit	Post-freeze
GB1	GB1 protein	15 runs	Cross-domain evidence	Supporting
E7b	CPEB3, Tetrahymena	120 trajectories	Sequential decisions	Frozen analysis

Constructive examples. For a graph consisting of a single edge, the residuals $e_A = (c, c)$ and $e_B = (c, -c)$ have identical pointwise RMSE, while $R_1(e_A) = 0$ and $R_1(e_B) = 2R_0$. Likewise, on a single mutation square, $e_A = (c, c, c, c)$ and $e_B = (c, -c, -c, c)$ have identical pointwise RMSE, while $R_2(e_A) = 0$ and $R_2(e_B) = 4R_0$. These examples show constructively that matching aggregate residual magnitude does not identify either first- or second-order relational fidelity.

A.2 Spectral form of residual geometry

Let

$$L_{\text{eval}} = U\Lambda U^\top, \qquad e = \sum_k \alpha_k u_k.$$

Then

$$\omega_e = \frac{\sum_k \lambda_k \alpha_k^2}{\sum_k \alpha_k^2}.$$

Thus residuals with similar Euclidean magnitude can have different Dirichlet energy, and therefore different first-order relational error, when their energy is allocated differently across graph-spectral directions.

Throughout this work, graph-Fourier frequency refers to Laplacian eigenvalue and is distinct from Walsh–Fourier interaction order. For the subsequent residual-spectrum audit, the constant mode is reported separately and nonconstant residual energy is partitioned into four fixed bands defined by fractions of $\lambda_{\max}$ rather than by equal-rank spectral quartiles. Because R_1 and the Laplacian Rayleigh quotient are algebraically linked when computed on common support, graph-spectral summaries that reduce to this same quadratic form are not counted as independent confirmations of first-order relational error.

B Experimental design, evaluation, and reproducibility

Table 1 summarizes the experimental components, evaluation roles, and inferential status used throughout the supplementary analyses.

B.1 Common normalization, structural diagnostics, and computational support

For the natural matched-risk benchmark, the Domingo mutation-square analysis, and the relation-aware-training study, the normalization scale is the population standard deviation of fitness in the corresponding training partition,

$$s_G = \mathrm{SD}\{f(v) : v \in V_{\mathrm{train}}\}.$$

The same s_G is used for pointwise, first-order, and second-order RMSE within a run. Consequently, $R_1 = c$ corresponds to a mutational-effect RMSE of $c\, s_G$, while $R_2 = c$ corresponds to a second-difference RMSE of $c\, s_G$. Likewise, a between-model difference $|\Delta R_1| = c$ represents a difference of $c\, s_G$ in mutational-effect RMSE on the same support.

These normalized quantities are not percentage errors in mutational or interaction effects because the denominator is the standard deviation of scalar fitness. Normalized risks are therefore interpreted jointly with structural diagnostics rather than used alone to assign biological importance. The Domingo mutation-square analysis reports interaction-sign and sign-epistasis diagnostics, GraphFLA provides landscape-level structural readouts, and the subsequent spectral audit reports scale-free allocations of full-landscape residual energy across fixed graph-frequency bands. These spectral energy fractions do not replace or renormalize the fixed R_2 endpoint.

For the sequential-decision analysis, relational quantities are computed retrospectively from fixed pre-decision surrogate snapshots and use the fixed standard deviation of fitness over the corresponding measured landscape. These truth-derived metrics are never supplied to the acquisition policy.

Evaluation cost scales linearly with declared support: $O(|V|)$ for pointwise quantities, $O(|E|)$ for first-order relations, and $O(|Q|)$ for mutation squares. Square support can therefore be fixed or sampled when exhaustive enumeration is impractical.

For the learned-direction matched-R_0 analysis, the normalization scale is the same frozen E1 training-partition scale s_G associated with each run. The held-out residual field is rescaled only to impose the declared exact normalized test risk $R_0 = \tau$. This calibration does not redefine the E1 training split or relational supports. Held-out first-order evaluation remains on test–test edges and held-out second-order evaluation remains on four-held-out mutation squares; full-landscape GraphFLA quantities are reported separately as reconstruction diagnostics.

B.2 General evaluation and inferential hierarchy

Pointwise RMSE and Pearson and Spearman correlations are reported alongside the relational endpoints. Interaction-sign recovery, epistasis classification, graph-Fourier residual-energy summaries, and GraphFLA descriptors are complementary structural diagnostics rather than additional algebraic measurements of the same quantity.

Repeated seeds and repeated splits quantify optimization, sampling, and resampling variability within systems but are not treated as independent biological evidence. Confirmatory cross-system inference therefore uses biological system as the independent unit. Algebraically related first-order summaries, including R_1 and Dirichlet-energy quantities computed on the same support, are not counted as independent confirmations.

For the subsequent formal inference on the Domingo mutation-square analysis, the ten seed–holdout configurations are treated as matched blocks across models. Paired mean differences are reported with 95% seed-cluster bootstrap intervals, resampling the five seeds while retaining both holdout modes within each seed. Configuration-level exact two-sided sign-flip tests are Holm-adjusted across the six combinations of three model contrasts and two structural endpoints. As a more conservative robustness analysis, the two holdout modes are first averaged within each seed and an exact two-sided sign-flip test is applied to the resulting five seed-level differences. Neither the ten configurations nor the five seeds are interpreted as independent biological replications.

The graph-Fourier extension is exploratory and uses the unchanged 30 full-landscape prediction fields and reports model means together with paired direction counts across the ten seed–holdout configurations and five seed-averaged contrasts. No additional hypothesis tests are attached to the spectral outcomes.

The learned-direction matched-R_0 analysis is likewise a secondary, non-confirmatory analysis. It uses the 84 fixed residual directions from the natural matched-risk benchmark and three fixed amplitudes as counterfactual reconstruction conditions. Seeds quantify within-system stability and the three amplitudes are conditions rather than replicates. Held-out labels are used only to impose exact test R_0 equality and to evaluate the already declared held-out relational endpoints; no model is retrained or reselected, and no new confirmatory inference is attached to this analysis.

The inferential status of each analysis is fixed throughout. The natural matched-risk benchmark is the frozen primary analysis. The Domingo mutation-square study is supporting: its formal paired inference was conducted after the analysis freeze on unchanged outputs, and its residual-spectrum audit is exploratory. The controlled matched-R_0 constructions are deterministic mechanism analyses added after the freeze, while the learned-direction matched-R_0 analysis is a magnitude-normalized counterfactual analysis of all four RNA systems. The relation-aware-training intervention is the prospectively frozen confirmatory analysis; its objective decomposition is a subsequent ablation, and the weighting and Pareto analyses are secondary audits of already selected validation-eligible objectives. StructRFM and the broader GraphFLA analyses were also added after the freeze. GB1 provides supporting cross-domain evidence, and the sequential-decision analysis evaluates downstream decisions. Outcome inspection is not used to alter model rosters, objective families, selection rules, matching margins, equivalence criteria, or evaluation endpoints.

B.3 External-model provenance and primary model-roster definition

The fixed primary roster contains One-hot Ridge, Pairwise Ridge, a Scratch Transformer, RNA-FM, RNAErnie, ERNIE-RNA, and RiNALMo. For each pretrained RNA model, the encoder is held fixed and a matched Ridge head is fitted to fitness labels. The benchmark therefore compares fixed representation quality under a common downstream protocol rather than maximum performance under model-specific fine-tuning.

The original pretraining corpora of the external RNA encoders are not fully auditable under a common provenance standard. We therefore do not interpret their relative performance as evidence about the effect of pretraining exposure itself, and no contamination-sensitive conclusion depends on assuming the absence of historical overlap.

Each model is evaluated on four RNA systems with three fixed seeds, yielding 84 fixed runs. Hyperparameters are selected from training and validation data only.

The frozen Claim-A comparison is One-hot Ridge versus the Scratch Transformer within biological system and seed. Qualifying pairs satisfy $|\Delta R_0| \leq 0.05$. The tolerance is a point-estimate matching rule and not a formal equivalence test. $R_{1,TT}$ is evaluated only on test–test edges. Other model-pair comparisons are supplementary and cannot modify the frozen Claim-A verdict.

StructRFM is evaluated after the primary-analysis freeze using the same four-system, three-seed protocol, producing 12 additional runs. These runs remain separate from the frozen seven-model roster.

B.4 Mutation-square support, assay-uncertainty procedure, and residual-spectrum audit

The Domingo mutation-square analysis contains three model classes, five fixed seeds, and two predefined holdout modes, yielding 30 runs. Within each seed–mode configuration, One-hot Ridge, Pairwise Ridge, and the Scratch Transformer use the same fixed train–test partition and exactly the same test-square support.

Across the ten seed–mode configurations, the primary second-order endpoint contains 13,438–14,907 mutation squares per run, and all 30 runs satisfy the predefined support-admissibility criterion. In the background-holdout mode, the support spans 832 backgrounds and all 45 represented site pairs. In the interaction-holdout mode, it spans 3,506–3,756 backgrounds across 9 site pairs. The stricter four-held-corner stratum contains 8,377–10,077 squares across configurations.

The two holdout modes probe different predefined regions of the combinatorial support and are not interpreted as exhaustive samples of every possible interaction context. Because models within a

seed–mode configuration are evaluated on identical square sets, between-model differences cannot be attributed to unequal evaluation support.

Alongside R_2, this analysis reports second-difference sign accuracy and sign-epistasis precision, recall, and F1. R_2 measures error in interaction magnitude, whereas the sign-based diagnostics measure interaction orientation and structural classification.

Where per-genotype measurement uncertainty is available, we propagate the available assay uncertainty to second differences and exclude mutation squares whose measured interaction is not distinguishable from uncertainty under the predefined $\alpha = 0.05$ criterion. The remaining support is termed assay-confident. This sensitivity is applied to sign recovery only; it does not redefine the frozen R_2 endpoint or its test-square support. Because uncertainty annotations are not uniformly available across all datasets, noise-aware quantities are not treated as universal cross-system endpoints.

The subsequent formal paired analysis uses only these unchanged outputs. For second-difference sign accuracy and sign-epistasis F1, all three paired model contrasts are evaluated across the ten identical seed–holdout blocks. No models are retrained or reselected and no evaluation support is changed.

A separate graph-Fourier audit, added after the analysis freeze, uses the full-landscape prediction field from each of the same 30 fixed runs. Residuals are defined over all 4,176 observed Domingo genotypes on the connected single-substitution graph with 26,336 edges. Let $L = D - A = U\Lambda U^\top$ be the combinatorial Laplacian and $e = U\alpha$ the residual expansion. The spectrum has smallest positive eigenvalue $\lambda_1 = 1.5924512103$ and $\lambda_{\max} = 23.7900251911$. The constant mode is reported separately. For the nonconstant residual energy, four fixed bands are defined by

$$
\begin{aligned}
Q_1 &: 0 < \lambda \le 0.25\lambda_{\max}, \\
Q_2 &: 0.25\lambda_{\max} < \lambda \le 0.50\lambda_{\max}, \\
Q_3 &: 0.50\lambda_{\max} < \lambda \le 0.75\lambda_{\max}, \\
Q_4 &: 0.75\lambda_{\max} < \lambda \le \lambda_{\max}.
\end{aligned}
$$

These bands contain 213, 1,560, 2,001, and 401 eigenvectors, respectively, with one additional constant eigenvector. For band Q_j, the reported nonconstant energy fraction is the sum of α_k^2 in that band divided by total nonconstant residual energy. The DC fraction is reported separately relative to total residual energy. Parseval reconstruction and the Laplacian-Rayleigh identity were verified numerically before summarization. The analysis is a full-landscape reconstruction diagnostic rather than a held-out-only frequency-generalization endpoint, and the bands are graph-Fourier frequency bands rather than Walsh interaction orders.

B.5 Foundation-model architecture, corpus provenance, and leakage audit

The relation-aware-training representation is a compact RNA encoder pretrained from scratch on an independently constructed, versioned RNACentral-derived corpus. Corpus construction and leakage controls are maintained separately from the external primary matched-risk benchmark.

An exhaustive post-window downstream-benchmark containment audit examined 94,108,635 corpus windows and detected no exact benchmark hits. A separate near-homology audit examined 10,000 sampled windows and detected no matches under the specified criterion. The latter analysis is explicitly a sampled audit and is not interpreted as exhaustive certification against every possible historical sequence relationship.

The encoder consists of a convolutional front end followed by an eight-layer Transformer with model dimension 256, eight attention heads, a 512-dimensional feed-forward block, rotary positional embeddings, and RMS normalization. Pretraining uses masked-language modeling with a 15% masking probability.

The production checkpoint is frozen before relation-aware fitting and before inspection of intervention outcomes. The same checkpoint is used for every intervention and control condition, so representation changes cannot be confounded with objective-function changes. Versioned corpus artifacts, training configuration, environment information, checkpoint identifiers, and audit records are retained in the supplementary reproducibility materials.

B.6 Relation-aware objective roster, validation eligibility, and inference

Relation-aware training is performed independently on CPEB3 and Tetrahymena using three predefined data-split replicates and two model seeds per split, giving six nested split–seed units per system. Within each unit, 16 candidates spanning 12 predefined objective families are fitted.

The roster includes conventional regression, regression-plus-ranking, generic contrastive and mutation-contrastive controls, an explicit graph-smoothing control, first-order preservation, joint first- and second-order preservation, and structured controls based on edge rewiring, target permutation, and potential-derived null construction. The graph-smoothing control provides an internal graph-aware comparator, whereas the relation-preservation objectives directly target the quantities evaluated by R_1 and R_2. Architecture, frozen representation, optimization schedule, prediction task, and training budget are otherwise fixed. All candidates are trained for 100 epochs without early stopping. All relation-preservation fitting targets are formed exclusively from training-labeled nodes and relations contained within the training support. Test–test edges and four-held-out mutation squares are reserved for held-out evaluation and are never used for fitting, validation selection, or model selection.

Candidate selection uses validation data only. A candidate is admissible when

$$R_0^{\mathrm{val}}(\mathrm{candidate}) \leq R_0^{\mathrm{val}}(\mathrm{baseline}) + 0.05.$$

Within each objective family and split–seed unit, the admissible candidate with the smallest predefined relational validation objective is selected, where

$$R_{\mathrm{rel}}^{\mathrm{val}} = R_1^{\mathrm{val}} + R_2^{\mathrm{val}}.$$

If no candidate in a family satisfies the pointwise validation constraint, the frozen fallback selection is retained for completeness but classified as exploratory-only and excluded from confirmatory inference.

Across the 144 selected family–unit bundles, 120 are confirmatory-eligible and 24 exploratory-only. All 24 validation-inadmissible selections occur in, CPEB3. Test outcomes are not used to alter candidate selection, eligibility, objective families, relational weighting, or the equivalence margin.

The complete campaign contains 96 fits per biological system, or 192 fits in total. Selected candidates are evaluated on held-out test partitions using the normalized endpoints defined in the main text.

Pointwise equivalence uses the prospectively specified ± 0.05 margin at $\alpha = 0.05$. Candidate–baseline differences are paired at the seed level, averaged within split, and then averaged within biological system. Only these system-level quantities enter the TOST procedure. Equivalence is established only when the corresponding 90% confidence interval lies completely inside $[-0.05, 0.05]$. Failure of this criterion indicates failure to establish equivalence and is not interpreted as evidence of pointwise harm.

B.7 GraphFLA descriptor methodology and GB1 supporting benchmark

GraphFLA is applied once to each of the four experimentally measured RNA landscapes and to each of the 84 corresponding frozen primary-benchmark predicted landscapes, yielding four true and 84 predicted baseline profiles. We evaluate local-optima ratio, gradient intensity, fitness autocorrelation, γ, epistasis-class proportions, and global-optimum accessibility. True and predicted analyses use identical observed genotype sets.

For each descriptor g, reconstruction error is

$$D_g = |g(\hat{f}) - g(f)|.$$

Global-optimum accessibility is not evaluated for F1*U because the observed single-mutation graph is fragmented. Local-optima and other descriptors that remain well defined on the fragmented support are retained.

The GraphFLA audit was added after the analysis freeze and is reported without confirmatory inference. It does not modify the primary benchmark endpoints, model selection, or confirmatory

inference. Its purpose is structural calibration: R_1 and R_2 quantify errors in local mutational effects and second differences, whereas GraphFLA measures changes in larger-scale reconstructed landscape properties.

The same frozen GraphFLA implementation is reused for the twelve controlled matched-R_0 landscapes: six on Domingo tRNA and six on Tetrahymena. These profiles form part of the deterministic matched-error mechanism analyses and are not pooled with the four true and 84 primary-benchmark prediction profiles.

The learned-direction matched-R_0 analysis also evaluates all $84 \times 3 = 252$ calibrated full-landscape reconstructions. These topology quantities are explicitly reconstruction diagnostics rather than held-out topology-generalization endpoints. The complete genotype-row support and retained GraphFLA vertex support are invariant under calibration. GraphFLA comparison edges, component counts, local optima, and accessibility can, however, change after calibration because exact neutral fitness ties in the measured landscape can be broken by the residual perturbation. Accordingly, comparison-edge and component counts are treated as structural outputs rather than support-equality gates. An implementation audit verified the expected true-landscape comparison-edge counts after removing exact neutral ties and found no input-row or retained-vertex support failures across the 252 reconstructions. Accessibility remains undefined for the fragmented F1*U landscape, leaving nine system–seed blocks for accessibility summaries at each amplitude.

The four-site GB1 protein landscape is evaluated separately using five model configurations and three fixed seeds, giving 15 runs. GB1 is supporting cross-domain biological evidence and is not treated as a confirmatory replication of the frozen primary effect-size claim.

B.8 Sequential-decision, permutation, and robustness analysis

The sequential-decision analysis compares a conventional regression baseline with a relational intervention on CPEB3 and Tetrahymena under matched initial observations, acquisition budget, start regime, and surrogate architecture. Each condition contains 30 paired replicates per system. Six observations initialize each trajectory, after which acquisition continues to a total budget of 20. Global greedy acquisition is the predefined primary policy and random acquisition is retained as a comparator. The intervention-level analysis treats biological system as the independent unit and uses a predefined minimum decision improvement of 0.05.

The separate incremental-association analysis contains all 120 trajectory-level rows: 60 from each biological system and 30 matched replicate identifiers across the baseline and intervention model roles. The reduced predictor block contains R_0, Spearman correlation, top-k recall, tail RMSE, and calibration error. The relational block adds $R_{1,TT}$ and $R_{2,4\mathrm{held}}$.

Significance is assessed using 2,000 Freedman–Lane permutations of reduced-model residuals within biological system. Uncertainty in partial R^2 is estimated using 2,000 system-aware bootstrap draws. Regression R^2 and partial R^2 are interpreted as explanatory associations within the evaluated setting rather than causal effects or out-of-sample predictive performance.

To assess sensitivity to predictor count and grouped dependence, an additional analysis uses leave-one-replicate-out prediction, with all rows sharing the held-out replicate identifier excluded during fitting. Reduced and relational models are additionally refitted with indicators for biological system and baseline-versus-intervention model role. This analysis is a robustness check and does not replace the frozen primary model.

All relational predictors are computed retrospectively from fixed pre-decision surrogate snapshots. At each snapshot, $R_{1,TT}$ uses only edges whose two endpoints remain unrevealed, and $R_{2,4\mathrm{held}}$ uses only mutation squares whose four corners remain unrevealed. The acquisition algorithm therefore never observes the truth-derived relational quantities subsequently used for evaluation.

C Post-freeze sensitivity analyses and additional results

Table 2 provides a compact summary of the principal empirical and controlled findings developed in the analyses below.

Table 2: Supplementary summary of the main empirical and controlled findings. Negative confirmatory results are retained under the predefined criteria.

Analysis	Comparison	Result	Conclusion
E1	$\Delta R_{1,TT} \geq 0.10$ at naturally matched R_0	max $\lvert\Delta R_{1,TT}\rvert = 0.0059$	Prespecified large first-order separation not established
E3	Similar R_2 vs. structural recovery	$R_2 \approx 1.31$; sign acc. $0.416/0.528/0.497$; lowest-band energy $0.262/0.208/0.515$	Different interaction and residual-frequency structure
Controlled matched-R_0	Exact R_0 equality on Domingo and Tetrahymena	Domingo $\Delta R_1 = 0.144/0.288/0.432$; Tetrahymena relational error separates at all three amplitudes	Deterministic mechanism: same pointwise error, different relational and reconstructed structure
Learned-direction matched-R_0	84 fixed learned residual directions rescaled to exact R_0	At $R_0 = 0.20$, max/min $R_{2,4\text{held}}$ reaches $2.82\times/1.50\times/1.42\times/1.17\times$ on CPEB3/Domingo/F1*U/Tetrahymena	Learned residual geometry remains distinct after pointwise-magnitude normalization
E5	Relational intervention	$\Delta R_{\text{rel}} = -0.0110/-0.0097$	Lower relational error in both systems
E5	R_0 equivalence, ± 0.05	0.0241; 90% CI $[-0.102, 0.150]$	Not established
E5 decomposition/Pareto	Objective-specific relational effects	Second-order-only $\Delta R_2 = -0.0033/-0.0064$; heterogeneous pointwise–relational trade-offs	Order-specific effects; no universal trade-off
E7b	Decision improvement ≥ 0.05	0.0434, $p = 0.124$	Not established
E7b	Incremental relational association	partial $R^2 = 0.756$; adjusted $R^2_{\text{partial}} = 0.058$	Strongly attenuated after system/role control

C.1 Alternative relational-error normalization

The primary analyses normalize pointwise and relational RMSEs by the standard deviation of training-set fitness,

$$s_G = \mathrm{SD}\{f(v) : v \in V_{\text{train}}\},$$

so that R_0, R_1, and R_2 are expressed relative to the overall fitness variation within each landscape. To test whether the relational conclusions depend on this denominator, we performed an additional sensitivity analysis using scales derived directly from the true relational quantities on the corresponding evaluation support.

For first-order error, we define

$$R_1^{(\Delta)} = \frac{\mathrm{RMSE}\left(\widehat{\Delta f} - \Delta f\right)}{\mathrm{SD}(\Delta f)}, \tag{10}$$

In Eq. (10), Δf denotes the true mutational effect on the exact edge support used by the corresponding endpoint. For second-order error,

$$R_2^{(\Delta^2)} = \frac{\mathrm{RMSE}\left(\widehat{\Delta^2 f} - \Delta^2 f\right)}{\mathrm{SD}(\Delta^2 f)}, \tag{11}$$

In Eq. (11), $\Delta^2 f$ denotes the true mutation-square second difference on the exact evaluated square support. Population standard deviations ($\mathrm{ddof} = 0$) are used throughout. As an additional orientation-robust check, we also replace each denominator by the RMS magnitude of the corresponding true relational quantity.

The analysis preserves the original experimental support and frozen decisions. No model was retrained, no matched-risk pair was redefined, no E5 candidate was reselected, and no validation-eligibility decision or confirmatory threshold was changed. For the primary matched-risk benchmark, the alternative first-order scale is computed on the fixed test–test edge support. For the mutation-square analysis, the second-order scale is computed on its dedicated fixed test-square set

Table 3: Post-freeze normalization sensitivity. "Effect-SD" normalizes by the population standard deviation of the corresponding true relational quantity on the exact evaluation support; "Effect-RMS" uses its RMS magnitude. The primary matched-risk benchmark reports the maximum first-order separation among the three fixed Claim-A matched-risk pairs. The mutation-square analysis reports model-level mean second-order error. The relation-aware intervention reports the system-level change for the joint first- and second-order preservation objective relative to regression.

Study	Quantity	Original	Effect-SD	Effect-RMS
E1	max matched $\lvert\Delta R_1\rvert$	0.005920	0.005659	0.005655
E3	One-hot Ridge R_2	1.313574	1.000338	1.000000
E3	Pairwise Ridge R_2	1.310930	0.998436	0.998099
E3	Scratch Transformer R_2	1.313679	1.000415	1.000077
E5	CPEB3 ΔR_{rel}	-0.011023	-0.014524	-0.014509
E5	Tetrahymena ΔR_{rel}	-0.009697	-0.007714	-0.007714

rather than on the broader diagnostic square support. For relation-aware training, the original selected candidate identities and validation-eligibility classifications are retained, and the alternative denominators are reconstructed separately for each biological system and split using the exact held-out support.

Table 3 summarizes the post-freeze normalization-sensitivity results under the alternative relational-error scales.

For the primary matched-risk benchmark, the maximum first-order separation across the three fixed One-hot-Ridge–Scratch-Transformer matched-risk pairs changed only from 0.005920 under the primary normalization to 0.005659 under Effect-SD and 0.005655 under Effect-RMS normalization. Thus the frozen conclusion that the predefined large matched-risk first-order separation was not established does not depend materially on the choice of normalization. The original predefined 0.10 threshold is not reapplied to the alternatively normalized values, because that threshold belongs to the frozen primary scale.

For the mutation-square analysis, alternative normalization leaves the aggregate second-order errors nearly indistinguishable across the three models. Mean Effect-SD-normalized R_2 values are 1.000338 for One-hot Ridge, 0.998436 for Pairwise Ridge, and 1.000415 for Scratch Transformer. The corresponding interaction-sign accuracies remain 0.416311, 0.527835, and 0.497181, respectively. Therefore, the finding that similar aggregate second-order error can coexist with different recovered interaction structure is not an artifact of normalizing by overall training-set fitness variation.

The relation-aware intervention result is also preserved. Under Effect-SD normalization, the joint first- and second-order preservation objective changes composite relational error by

$$\Delta R_{\mathrm{rel}} = -0.014524 \quad \text{on CPEB3}$$

and

$$\Delta R_{\mathrm{rel}} = -0.007714 \quad \text{on Tetrahymena.}$$

The corresponding Effect-RMS values are -0.014509 and -0.007714, respectively. Thus the direction of the relational intervention is unchanged in both biological systems under either alternative scale.

The componentwise Effect-SD changes further show that the two systems need not improve through identical relational mechanisms:

$$\begin{aligned} \text{CPEB3:} \quad & \Delta R_1 = -0.006291, \quad \Delta R_2 = -0.008233, \\ \text{Tetrahymena:} \quad & \Delta R_1 = +0.000006, \quad \Delta R_2 = -0.007720. \end{aligned}$$

CPEB3 therefore shows reductions in both first- and second-order relational error, whereas the Tetrahymena change is almost entirely second order.

As an implementation check, all six relation-aware biological-system–split support reconstructions matched the fixed split hashes and support counts, and the stored raw RMSE numerators reproduced the original normalized metrics before the alternative denominators were applied. The sensitivity analysis retains all 144 fixed family-level selections, including the original 120 confirmatory-eligible

selections. The seven eligible paired joint-preservation units are nested split/seed stability measurements and are not treated as seven independent biological replicates; the independent relation-aware evidence remains the two biological systems.

C.2 MATCHED-RISK AND MODEL-ROSTER SENSITIVITIES

As an additional sensitivity analysis, we broadened the pairing set from the frozen One-hot Ridge–Scratch Transformer comparison to all seven primary-benchmark model families while imposing tighter pointwise matching.

At $|\Delta R_0| \leq 0.005$, 19 same-system, same-seed pairs were available across all four RNA systems. Their maximum and mean $|\Delta R_{1,TT}|$ were 0.0670 and 0.0199, respectively. At $|\Delta R_0| \leq 0.01$, 40 pairs were available across all four systems, with maximum and mean relational separations of 0.0869 and 0.0222. Neither sensitivity analysis produced a first-order separation reaching the frozen 0.10 target. Because the model-pair universe was broadened after the primary analysis, these comparisons are supplementary and do not redefine the predefined Claim-A test.

Because $R_{1,TT}$ is normalized by the landscape fitness scale s_G, a difference such as 0.00592 represents $0.00592\, s_G$ in held-out mutational-effect RMSE on the same evaluation support. Such values are not percentages of mutational effects, and their magnitude alone is not interpreted as evidence of biologically important structural divergence.

Descriptive comparisons within the broader primary benchmark nevertheless show that closely matched pointwise errors need not yield numerically identical first-order relational errors. On CPEB3, RNAErnie and RiNALMo differ by 0.00268 in mean R_0 while differing by 0.01124 in mean $R_{1,TT}$. On Domingo tRNA, RNA-FM and RNAErnie differ by 0.00471 in mean R_0 and 0.01010 in mean $R_{1,TT}$. These differences are modest on the normalized scale and are reported as exploratory.

The subsequent StructRFM analysis produced valid $R_{1,TT}$ measurements across all 12 additional four-system, three-seed runs. These runs remain separate from the fixed primary roster and do not modify the primary matched-risk conclusion.

C.3 MUTATION-SQUARE SUPPORT AUDIT, ASSAY-NOISE SENSITIVITY, AND FORMAL PAIRED INFERENCE

Within every mutation-square-analysis seed and holdout mode, One-hot Ridge, Pairwise Ridge, and the Scratch Transformer are evaluated on exactly the same frozen mutation-square support. All 30 mutation-square-analysis runs satisfy the predefined second-order support-admissibility criterion.

The primary support contains 13,438–14,907 test squares per configuration. Background-holdout supports span 832 backgrounds and all 45 represented site pairs, whereas interaction-holdout supports span 3,506–3,756 backgrounds across 9 site pairs. The stricter four-held-corner strata contain 8,377–10,077 squares across configurations. Consequently, between-model differences in R_2 or sign recovery within a configuration cannot arise from evaluating the models on different square sets. Conclusions nevertheless remain conditional on these predefined support regimes.

Under the assay-noise sensitivity analysis, approximately 62.5% of the mutation-square-analysis mutation squares are classified as assay-ambiguous under the predefined uncertainty criterion, leaving on average approximately 5.3×10^3 assay-confident squares per run. Restricting sign recovery to this support yields mean second-difference sign accuracies of 0.418 for One-hot Ridge, 0.540 for Pairwise Ridge, and 0.494 for the Scratch Transformer, preserving the ordering observed on the full fixed support.

For reference, mean primary R_2 values are 1.313574, 1.310930, and 1.313679 for One-hot Ridge, Pairwise Ridge, and the Scratch Transformer, respectively. The One-hot Ridge–Scratch Transformer difference is only approximately 1.05×10^{-4}, whereas their mean sign accuracies are approximately 0.416 and 0.497. Pairwise Ridge achieves mean sign-epistasis F1 of approximately 0.285, compared with 0 for One-hot Ridge and near zero for the Scratch Transformer.

Table 4: Post-freeze formal paired inference for E3 structural diagnostics. Δ is the paired difference in the direction indicated. Confidence intervals are 95% seed-cluster bootstrap intervals. p_{cfg} denotes the exact two-sided sign-flip test over the ten matched seed–holdout configurations, p_{Holm} is adjusted across the six model-by-endpoint comparisons, and p_{seed} is the conservative exact test after averaging the two holdout modes within each of the five seeds.

Endpoint	Contrast	Δ	95% CI	$\mathbf{p}_{\mathrm{Holm}}$	$\mathbf{p}_{\mathrm{seed}}$
Sign acc.	Pairwise–One-hot	$+0.1115$	$[0.0977, 0.1253]$	0.0117	0.0625
Sign acc.	Scratch–One-hot	$+0.0809$	$[0.0725, 0.0877]$	0.0117	0.0625
Sign acc.	Pairwise–Scratch	$+0.0307$	$[0.0167, 0.0446]$	0.0117	0.0625
Sign-epi. F1	Pairwise–One-hot	$+0.2855$	$[0.2612, 0.3098]$	0.0117	0.0625
Sign-epi. F1	Scratch–One-hot	$+0.0081$	$[0.0000, 0.0237]$	0.5000	0.5000
Sign-epi. F1	Pairwise–Scratch	$+0.2774$	$[0.2397, 0.3095]$	0.0117	0.0625

These diagnostics address different aspects of reconstruction: R_2 quantifies second-difference magnitude error, whereas sign accuracy and sign-epistasis F1 characterize interaction orientation and structural class.

Post-freeze formal paired inference. We quantified the consistency of the sign-based differences using the frozen 30-run mutation-square-analysis output matrix only. For each structural endpoint, all three model contrasts were computed within the ten identical seed–holdout configurations. The resulting six comparisons were evaluated using exact two-sided sign-flip tests, with Holm adjustment across the complete six-comparison family. Confidence intervals were obtained by resampling the five seeds as clusters, with both holdout modes retained within each seed.

Pairwise Ridge exceeded One-hot Ridge in sign accuracy in all ten seed–holdout configurations, as did the Scratch Transformer. Pairwise Ridge exceeded the Scratch Transformer in nine of ten configurations. After averaging the two holdout modes within each seed, all three sign-accuracy contrasts retained the same direction in all five seeds.

For sign-epistasis F1, Pairwise Ridge exceeded both One-hot Ridge and the Scratch Transformer in all ten configurations and in all five seed averages. The Scratch–One-hot contrast was qualitatively different: eight of ten configuration-level differences were ties, and only two of the five seed-averaged differences were positive.

The five consistent directional contrasts remained significant at the configuration level after Holm correction ($p_{\mathrm{Holm}} = 0.011719$). At the stricter seed-cluster level, each of these contrasts produced an exact two-sided $p = 0.0625$. With five nonzero seed-level differences, $0.0625 = 2/2^5$ is the minimum attainable two-sided exact sign-flip value. The seed-level result is therefore reported as a conservative small-n robustness analysis rather than interpreted as failure of the observed directional consistency. These analyses establish within-Domingo stability only and do not constitute independent biological replication.

C.4 Residual spectrum and structural consequences on reconstructed landscape topology

The graph-Fourier audit first asks whether the nearly identical aggregate second-order errors correspond to similar full-landscape residual-frequency organization. They do not. Numerical integrity checks gave maximum relative Parseval error 7.07×10^{-15} and maximum Rayleigh-identity difference 7.64×10^{-14} across the frozen prediction fields.

Table 5 reports the corresponding model-level graph-Fourier residual-energy summaries.

The One-hot Ridge–Scratch Transformer comparison is especially informative because their mean primary R_2 values are nearly identical (1.313574 versus 1.313679) while their residual spectra differ strongly. Scratch places 0.253569 more nonconstant residual energy in Q_1 than One-hot Ridge. This difference is positive in all ten seed–holdout configurations and all five seed averages. Scratch correspondingly places less energy in Q_2, Q_3, and Q_4, with mean differences of -0.127522, -0.107406, and -0.018641, respectively; each direction holds in all ten configurations and all five seed averages. Its DC fraction is also larger by 0.073521 on average, positive in nine of ten configurations and all five seed averages.

Table 5: E3 model-level graph-Fourier residual-energy summaries. DC is the constant-mode fraction of total residual energy. Q_1–Q_4 are fractions of nonconstant residual energy in the four fixed $\lambda_{\max}$-relative graph-frequency bands.

Model	DC	$\mathbf{Q_1}$	$\mathbf{Q_2}$	$\mathbf{Q_3}$	$\mathbf{Q_4}$	$\mathbf{Q_4 - Q_1}$
One-hot Ridge	0.011837	0.261712	0.379655	0.305635	0.052999	-0.208713
Pairwise Ridge	0.014612	0.207576	0.400063	0.334277	0.058085	-0.149491
Scratch Transformer	0.085358	0.515281	0.252132	0.198229	0.034358	-0.480923

Table 6: E3 model-level means for aggregate second-order error, interaction-sign recovery, and selected GraphFLA descriptor distortions. Structural quantities are absolute deviations from the experimentally observed Domingo tRNA landscape.

Model	$\mathbf{R_2}$	Sign acc.	$\mathbf{D}_{\mathrm{LO}}$	$\mathbf{D}_{\mathrm{acc}}$	$\mathbf{D}_{\mathrm{grad}}$
One-hot Ridge	1.313574	0.416311	0.020354	0.162596	0.271867
Pairwise Ridge	1.310930	0.527835	0.019684	0.146025	0.204347
Scratch Transformer	1.313679	0.497181	0.018702	0.146432	0.334790

Pairwise Ridge occupies a different spectral regime again. Relative to One-hot Ridge, Pairwise has 0.054136 less Q_1 energy and 0.020408, 0.028642, and 0.005086 more energy in Q_2, Q_3, and Q_4, respectively. Each of these directions occurs in nine of ten configurations and all five seed averages. Relative to Scratch, Pairwise has 0.307705 less Q_1 energy and more energy in each higher band ($+0.147930$, $+0.136048$, and $+0.023727$ for Q_2–Q_4); all four directions hold in all ten configurations and all five seed averages.

These spectral comparisons are exploratory; no hypothesis tests are attached to the spectral outcomes, and the repeated configurations quantify within-Domingo stability rather than independent biological replication. They demonstrate that nearly identical aggregate second-order error can arise from learned models whose residuals occupy markedly different graph-frequency regions. The bands should not be interpreted as Walsh interaction-order components.

We next test whether the interaction-sign and residual-spectrum differences correspond to larger-scale changes in reconstructed landscape structure using the complete fixed mutation-square-analysis prediction set. The analysis contains one experimentally observed Domingo tRNA landscape and all 30 predicted landscapes from the mutation-square analysis, corresponding to three model classes, five seeds, and two holdout modes. No model is retrained or reselected.

All 31 landscapes are reconstructed on the identical observed genotype set, containing 4,176 genotypes and 26,336 single-mutation edges, and the resulting graph is connected in every case. Before inspecting the GraphFLA outcomes, local-optima ratio and global-optimum accessibility were declared as the primary structural endpoints. Gradient intensity, autocorrelation, γ, and GraphFLA epistasis-class proportions were retained as secondary descriptors.

For a structural descriptor g, distortion relative to the experimentally observed landscape is defined as

$$D_g = \left| g(\widehat{f}) - g(f) \right|, \tag{12}$$

In Eq. (12), f is the measured Domingo tRNA landscape and $\widehat{f}$ is the corresponding fixed mutation-square-analysis prediction. The observed landscape has local-optima ratio 0.02059 and global-optimum accessibility 0.83740.

Table 6 reports the model-level means for second-order error, interaction-sign recovery, and the selected reconstruction descriptors.

The One-hot Ridge–Scratch Transformer comparison provides the closest connection to the main second-order result. Their mean aggregate second-order errors are nearly identical,

$$R_2 = 1.313574 \quad \text{and} \quad R_2 = 1.313679,$$

respectively, giving a mean difference of only 1.05×10^{-4}. Across the 10 matched seed–holdout configurations, this aggregate R_2 difference is not consistently directional: Scratch Transformer has lower R_2 in three pairs and higher R_2 in seven.

Table 7: Matched One-hot Ridge–Scratch Transformer comparison across the 10 identical seed–holdout configurations. Negative structural differences indicate lower distortion for Scratch Transformer. Exact sign-flip tests quantify within-Domingo consistency and are not treated as independent biological replication.

Quantity	**One-hot**	**Scratch**	**Δ Scratch–One-hot**	p_{flip}
R_2	1.313574	1.313679	+0.000105	0.210938
Interaction-sign accuracy	0.416311	0.497181	+0.080871	0.001953
D_{LO}	0.020354	0.018702	-0.001652	0.001953
D_{acc}	0.162596	0.146432	-0.016164	0.001953
D_{grad}	0.271867	0.334790	+0.062923	0.001953

In contrast, Scratch Transformer has higher interaction-sign accuracy in all 10 matched configurations. Its mean sign accuracy is 0.497181, compared with 0.416311 for One-hot Ridge, corresponding to a mean paired increase of 0.080871. An exact two-sided sign-flip test over the 10 matched configurations gives $p = 0.001953$. The broader formal paired analysis, including multiplicity adjustment and seed-cluster robustness, is reported in Table 4.

The same paired direction is observed for both predefined primary structural endpoints. Local-optima-ratio distortion decreases from 0.020354 for One-hot Ridge to 0.018702 for Scratch Transformer, a paired mean change of -0.001652. Global-optimum-accessibility distortion decreases from 0.162596 to 0.146432, a paired mean change of -0.016164. Scratch Transformer has lower distortion for both endpoints in all 10 matched configurations; the exact two-sided sign-flip value is $p = 0.001953$ for each comparison.

Table 7 summarizes the matched One-hot Ridge–Scratch Transformer structural comparison across the ten seed–holdout configurations.

The structural differences are not uniformly aligned across all descriptors. In particular, Scratch Transformer has larger gradient-intensity distortion than One-hot Ridge in all 10 matched configurations (0.334790 versus 0.271867; paired mean difference $+0.062923$). GraphFLA sign-epistasis distortion differs only slightly between the two models (0.350849 versus 0.353909), while reciprocal-sign-epistasis distortion is identical at 0.182046. Thus, improved interaction-sign recovery does not imply uniformly improved reconstruction of every landscape descriptor.

Across all 30 mutation-square-analysis prediction runs, interaction-sign accuracy is negatively associated with distortion of the two predefined structural endpoints: Spearman $\rho = -0.592$ for local-optima-ratio distortion and $\rho = -0.694$ for global-optimum-accessibility distortion. These pooled associations are exploratory. Within-model associations are weaker and are not statistically distinguishable from zero, while the One-hot Ridge correlations are undefined because its reconstructed primary topology values are constant across the corresponding runs. We therefore place greater weight on the matched model comparisons than on the pooled correlations.

Overall, the mutation-square analyses show that nearly identical aggregate second-order error can coexist with systematically different interaction-sign recovery, graph-frequency residual organization, and reconstruction of local-optimum and accessibility structure. At the same time, the lack of uniform improvement across all GraphFLA descriptors indicates that relational fidelity is structure-specific rather than a single scalar ordering of reconstructed landscape quality. Because all 30 predictions concern one biological landscape, these results establish within-Domingo robustness rather than independent cross-system replication.

C.5 Additional GraphFLA reconstruction results

GraphFLA is applied to the four experimentally observed RNA landscapes and all 84 fixed primary-benchmark predictions, yielding 88 landscape profiles. Descriptor reconstruction errors show heterogeneous relationships with R_0 and $R_{1,TT}$ rather than reducing to either metric alone.

Gradient-intensity distortion shows the strongest pooled association, with Spearman correlations of 0.813 with R_0 and 0.955 with $R_{1,TT}$. Other descriptors, including sign epistasis, accessibility, fitness autocorrelation, and local-optima structure, show weaker or qualitatively different relationships.

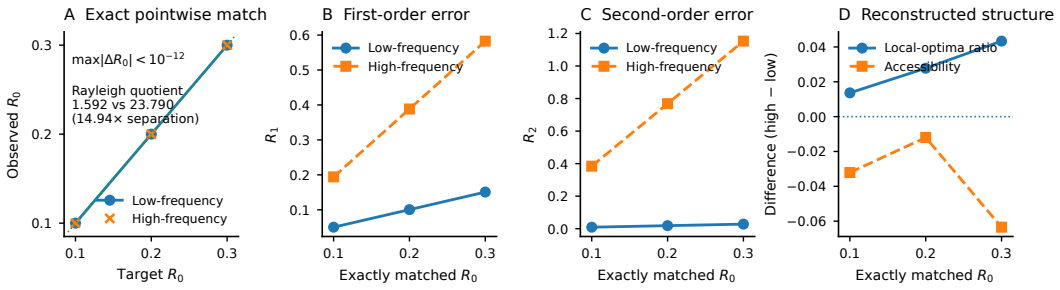


Figure 4: **Controlled exact matched-R_0 residual-geometry mechanism on Domingo tRNA.** (A) Smooth/low-frequency and alternating/high-frequency residual directions are scaled to the same pointwise risk at each amplitude; their Laplacian Rayleigh quotients differ by approximately $14.94\times$. (B) First-order relational error R_1 separates despite exact R_0 matching. (C) Second-order relational error R_2 separates more strongly. (D) The reconstructed local-optima ratio and global-optimum accessibility also differ, showing that identical aggregate pointwise error does not uniquely determine reconstructed landscape structure. The construction is a controlled mechanism demonstration, not an estimate of naturally occurring model prevalence.

A particularly close comparison occurs for Tetrahymena ERNIE-RNA and One-hot Ridge, whose R_0 values differ by approximately 0.00154 while their reconstructed landscapes differ across several GraphFLA descriptors. This comparison is illustrative rather than a confirmatory matched-risk test.

Together with the mutation-square analysis, these results provide structural calibration for the relational framework: aggregate relational RMSE quantifies error magnitude, whereas interaction signs, epistatic classes, graph-frequency residual organization, and landscape descriptors characterize complementary aspects of the structure induced by those errors.

C.6 Controlled matched-R_0 residual-geometry constructions

To isolate residual organization from aggregate pointwise magnitude, we performed deterministic matched-R_0 constructions after the analysis freeze on the complete observed Domingo tRNA and Tetrahymena landscapes. In both cases the residual directions are fixed independently of the subsequent relational and GraphFLA outcomes, normalized to unit RMS over the complete observed genotype set, and scaled to target $R_0 \in \{0.10, 0.20, 0.30\}$. Within each pair, R_0 is exactly equal by construction rather than estimated as statistically equivalent.

Figure 4 illustrates the controlled exact matched-R_0 residual-geometry construction on Domingo tRNA.

Domingo tRNA. The observed Domingo genotype graph contains 4,176 vertices, 26,336 single-mutation edges, and 70,713 mutation squares and forms a single connected component. Let L be its combinatorial Laplacian. The constant mode is excluded. The smooth residual direction is selected from the eigenspace associated with the smallest positive Laplacian eigenvalue, while the alternating residual direction is selected from the eigenspace associated with the largest eigenvalue. The corresponding eigenvalues are

$$\lambda_{\mathrm{lo}} = 1.592451, \qquad \lambda_{\mathrm{hi}} = 23.790025,$$

giving a Rayleigh-quotient ratio of approximately

$$\frac{\lambda_{\mathrm{hi}}}{\lambda_{\mathrm{lo}}} = 14.939.$$

Both relevant eigenspaces are one-dimensional for this construction. Residual directions are centered, deterministically oriented, and normalized to unit RMS before scaling.

Using the population standard deviation of fitness over the complete observed Domingo landscape as the fixed construction scale, we generate smooth and alternating surrogate landscapes at the three target pointwise-error levels.

Table 8: Relational errors for the Domingo controlled matched-R_0 construction. Pointwise error is exactly matched within each smooth–alternating pair.

$\mathbf{R_0}$	Residual	$\mathbf{R_0}$	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{R_{rel}}$
0.10	Smooth	0.1000	0.0503	0.0094	0.0596
0.10	Alternating	0.1000	0.1942	0.3843	0.5786
0.20	Smooth	0.2000	0.1005	0.0188	0.1193
0.20	Alternating	0.2000	0.3884	0.7687	1.1571
0.30	Smooth	0.3000	0.1508	0.0282	0.1789
0.30	Alternating	0.3000	0.5827	1.1530	1.7357

Table 9: Primary GraphFLA endpoints for the Domingo controlled matched-R_0 construction.

$\mathbf{R_0}$	Residual	Local-optima ratio	Global-optimum accessibility
0.10	Smooth	0.016523	0.864943
0.10	Alternating	0.030172	0.832854
0.20	Smooth	0.013649	0.861590
0.20	Alternating	0.041427	0.849617
0.30	Smooth	0.012931	0.856561
0.30	Alternating	0.056274	0.793103

Table 8 reports the relational errors for the Domingo controlled matched-R_0 construction.

The alternating-minus-smooth first-order differences are 0.1440, 0.2879, and 0.4319 at $R_0 = 0.10$, 0.20, and 0.30, respectively. The corresponding second-order differences are 0.3749, 0.7499, and 1.1248, and the composite relational differences are 0.5189, 1.0378, and 1.5567. The predefined first-order direction is therefore satisfied at all three amplitudes.

The first-order result is a direct consequence of the graph-spectral construction and should not be interpreted as evidence independent of the Rayleigh-quotient contrast. It provides an explicit realization of the identity derived in Appendix A.1: fixed residual norm does not fix Dirichlet energy.

We next evaluate whether the matched-error residuals reconstruct different landscape-level structure using the same fixed GraphFLA implementation as the other supplementary structural analyses.

Table 9 reports the corresponding GraphFLA reconstruction endpoints for Domingo.

The alternating-minus-smooth local-optima-ratio differences are $+0.013649$, $+0.027778$, and $+0.043343$ as R_0 increases from 0.10 to 0.30. Global-optimum-accessibility differences are -0.032088, -0.011973, and -0.063458, respectively. Thus the reconstructed topology differs at every exactly matched pointwise-error level.

The experimentally observed Domingo landscape has local-optima ratio 0.020594 and global-optimum accessibility 0.837404. The smooth construction is closer to the observed local-optima ratio at all three amplitudes. Accessibility is descriptor-specific: the alternating construction is closer to the observed accessibility at $R_0 = 0.10$ and 0.20, whereas the smooth construction is closer at $R_0 = 0.30$. Accordingly, the controlled result is not interpreted as showing that high-frequency residuals are universally less faithful; it shows that equal pointwise error does not uniquely determine reconstructed structure.

Secondary descriptors show the same structure-specific divergence. The alternating-minus-smooth autocorrelation differences are -0.0463, -0.1154, and -0.2016 across the three amplitudes, while the corresponding γ differences are -0.0716, -0.2086, and -0.3695. Reciprocal-sign epistasis differences increase from $+0.0374$ to $+0.0961$ and $+0.1551$. Other epistasis classes show different directional patterns.

Tetrahymena extension. We repeat the fixed matched-error mechanism on the complete observed Tetrahymena landscape, whose connected single-substitution graph contains 64,020 measured genotypes and 752,086 edges. The extension uses fixed smooth/low-frequency and alternating/high-frequency residual directions under the frozen extension protocol. At unit normalized residual amplitude, the smooth and alternating directions have first-order factors $R_1/R_0 = 0.57017$ and 1.65044, respectively, and second-order factors $R_2/R_0 = 0.05258$ and 2.74559.

Table 10: Relational errors for the Tetrahymena matched-R_0 extension. Pointwise error is exactly matched within each smooth–alternating pair.

$\mathbf{R_0}$	**Residual**	$\mathbf{R_0}$	$\mathbf{R_1}$	$\mathbf{R_2}$	$\mathbf{R_{rel}}$
0.10	Smooth	0.1000	0.057017	0.005258	0.062275
0.10	Alternating	0.1000	0.165044	0.274559	0.439603
0.20	Smooth	0.2000	0.114034	0.010516	0.124550
0.20	Alternating	0.2000	0.330088	0.549118	0.879206
0.30	Smooth	0.3000	0.171051	0.015774	0.186825
0.30	Alternating	0.3000	0.495132	0.823677	1.318809

Table 11: Primary GraphFLA endpoints for the Tetrahymena matched-R_0 extension.

$\mathbf{R_0}$	**Residual**	**Local-optima ratio**	**Global-optimum accessibility**
0.10	Smooth	0.016354	0.851172
0.10	Alternating	0.018541	0.870822
0.20	Smooth	0.016042	0.897095
0.20	Alternating	0.022571	0.823477
0.30	Smooth	0.015417	0.916838
0.30	Alternating	0.026804	0.821197

Table 10 reports the relational errors for the Tetrahymena matched-R_0 extension.

The alternating/high-frequency residual therefore produces larger first- and second-order relational error at every exactly matched R_0. The corresponding alternating-minus-smooth differences in R_{rel} are 0.377328, 0.754656, and 1.131984.

The same separation extends to reconstructed topology.

Table 11 reports the corresponding GraphFLA reconstruction endpoints for Tetrahymena.

The experimentally observed Tetrahymena landscape has local-optima ratio 0.0154483 and global-optimum accessibility 0.8773196. The smooth construction is closer to the observed local-optima ratio at all three amplitudes. Accessibility closeness is mixed: the alternating construction is closer at $R_0 = 0.10$, whereas the smooth construction is closer at $R_0 = 0.20$ and 0.30. As on Domingo, the result is therefore a geometry-dependent change in reconstructed structure rather than a universal ordering in which one residual family is closer for every descriptor.

Taken together, the Domingo and Tetrahymena constructions provide a deterministic existence-and-portability demonstration on two empirical RNA landscapes. They do not estimate biological prevalence and are not treated as two confirmatory biological replications. The three amplitudes within each landscape are experimental conditions rather than independent replicates, and no p-values are attached to the campaign. The conclusion is limited to the mechanism established by construction: identical aggregate pointwise error can correspond to different relational geometry and different reconstructed landscape topology.

C.7 Relation-aware control families and validation eligibility

The relation-aware control analyses further distinguish relational improvement from pointwise equivalence. Three objective families satisfied the predefined pointwise TOST criterion, but none reduced relational error in both biological systems. Conversely, joint first- and second-order preservation reduced relational error in both systems but did not satisfy the pointwise-equivalence criterion.

The explicit graph-smoothing control produced $\Delta R_{\mathrm{rel}} = +0.5214$ on CPEB3 and -0.0373 on Tetrahymena. Its mean pointwise difference was -0.0180, with a 90% confidence interval of approximately $[-0.0510, 0.0149]$, so its pointwise-equivalence criterion was also not established. The two-system relational reduction observed for joint first- and second-order preservation was therefore not reproduced by generic graph smoothing alone.

Validation eligibility differed between systems. All 72 selected Tetrahymena family–unit bundles were confirmatory-eligible, whereas 48 of 72 CPEB3 bundles were eligible. Validation-inadmissible

Table 12: Post-freeze E5 objective decomposition. Differences are held-out changes relative to conventional regression for the second-order-only arm. Negative values indicate lower relational error.

System	ΔR_1	ΔR_2	ΔR_{rel}
CPEB3	+0.007448	-0.003279	+0.004169
Tetrahymena	-0.001541	-0.006355	-0.007896

fallback selections were retained as exploratory outputs and excluded from confirmatory equivalence inference.

The primary joint-preservation result remains $\Delta R_{\mathrm{rel}} = -0.0110$ on CPEB3 and -0.0097 on Tetrahymena. The corresponding mean system-level pointwise difference is $+0.0241$, with a hierarchical TOST 90% confidence interval of $[-0.1022, 0.1505]$. Pointwise equivalence is therefore not established.

C.8 Objective decomposition and Pareto audit

To distinguish first- and second-order contributions to the relation-aware intervention, we performed a subsequent ablation adding a true second-order-only objective. The original 192-fit campaign and its fixed selections remain unchanged.

The added objective sets $\lambda_1 = 0$ and evaluates

$$\lambda_2 \in \{0.03, 0.10, 0.30\}$$

under the original fixed encoder, architecture, optimization schedule, training budget, biological-system splits, and model seeds. Three candidates are fitted for each of the 12 biological-system–split–seed units, yielding 36 new fits.

Candidate selection is validation-only and follows the original relation-aware-training rule. Second-order-only candidates must satisfy

$$R_0^{\mathrm{val}}(\mathrm{candidate}) \leq R_0^{\mathrm{val}}(\mathrm{baseline}) + 0.05,$$

after which the admissible candidate with the smallest original validation relational objective is selected. Held-out outcomes are not used to select λ_2.

The second-order-only family is validation-admissible in 10 of 12 split–seed units: four of six in CPEB3 and all six in Tetrahymena. Consequently, comparisons involving non-admissible CPEB3 units are not promoted to confirmatory evidence.

Table 12 summarizes the post-freeze E5 objective decomposition relative to conventional regression.

Second-order-only preservation reduces R_2 in both systems (-0.003279 in CPEB3 and -0.006355 in Tetrahymena), but its effect on R_1 differs by system. In CPEB3, the first-order increase is sufficiently large that the composite relational change is positive ($\Delta R_{\mathrm{rel}} = +0.004169$). In Tetrahymena, both relational orders move downward and the composite change is $\Delta R_{\mathrm{rel}} = -0.007896$.

Across the two biological systems, the mean second-order-only versus regression composite change is -0.001864, with a system-level interval of approximately $[-0.039953, 0.036225]$. This cross-system summary is therefore not interpreted as establishing a uniform composite improvement from the second-order-only objective.

Tetrahymena provides the clearest direct decomposition because the relevant preservation families are validation-admissible across all six units. Relative to regression,

$$\Delta R_{\mathrm{rel}} = -0.008941$$

for first-order preservation,

$$\Delta R_{\mathrm{rel}} = -0.007896$$

for second-order-only preservation, and

$$\Delta R_{\mathrm{rel}} = -0.009697$$

for joint first- and second-order preservation. Relative to first-order preservation, the joint objective changes composite relational error by only -0.000756; relative to second-order-only preservation, the joint change is -0.001800.

For CPEB3, the joint objective changes R_{rel} by $+0.009342$ relative to second-order-only preservation on the single common validation-admissible unit. Because only one common unit contributes to this direct contrast, it is reported as a secondary result and is not treated as evidence of a system-level ordering between those two objectives.

These results indicate that first- and second-order preservation are not simply interchangeable regularizers. The second-order term preferentially changes second-order fidelity, while the combined objective can redistribute error across relational orders. The resulting trade-off is system dependent, and improvement in R_2 alone does not imply improvement in R_{rel}.

Post-freeze weighting sensitivity. The primary relation-aware composite uses the operational convention

$$R_{\mathrm{rel}} = R_1 + R_2.$$

Because first- and second-order errors need not be biologically commensurate, we additionally evaluate the selected predictions under

$$R_{\mathrm{rel}}(\alpha) = \alpha R_1 + (1-\alpha) R_2, \qquad 0 \le \alpha \le 1.$$

This is a re-expression of already selected held-out predictions, performed after the analysis freeze: no model is retrained, no candidate is reselected, no validation-eligibility classification is changed, and no new confirmatory inference is performed. On the common validation-eligible comparison set, the joint-preservation contrast relative to conventional regression remains negative across the full weighting sweep on CPEB3. On Tetrahymena it remains negative until the weight is effectively all first order ($\alpha \approx 0.999$), consistent with the very small first-order contrast in that system. Other preservation arms show greater weight sensitivity. The primary equal-weight result is therefore not presented as a unique biological scalarization; R_1 and R_2 remain reported separately throughout.

Post-freeze Pareto audit. We also perform a Pareto audit over the validation-eligible conventional-regression, first-order-only, second-order-only, and joint-preservation selections. For selected objective j within split–seed unit u, we represent held-out change relative to the corresponding regression reference as

$$z_{u,j} = \left(\Delta R_{0,u,j}, \Delta R_{\mathrm{rel},u,j}\right),$$

where smaller values are preferred on both coordinates. Within a unit, selection a dominates selection b if it is no worse on both coordinates and strictly better on at least one; the Pareto set contains the non-dominated selections.

This audit is applied only after the original validation-based selections are fixed. Held-out outcomes are not used to change objective families, hyperparameters, admissibility, or candidate identity, and Pareto membership is not used as a new selection rule. The analysis is therefore secondary rather than confirmatory.

The held-out pointwise–relational trade-offs are heterogeneous across split–seed units and objective families. In some units, a relation-preserving selection lowers R_{rel} without a corresponding increase in R_0, showing that a relational gain need not require pointwise degradation in every fitted unit. In other units the two coordinates trade against one another, and no single preservation family uniformly dominates across the evaluated systems. The Pareto audit therefore characterizes the observed trade-off structure but does not alter the predefined system-level equivalence result: formal pointwise equivalence for the confirmatory intervention remains unestablished.

C.9 Downstream association and robustness analyses

The predefined incremental-association analysis contains 120 trajectory-level observations, with 60 from each biological system. The reduced conventional model containing R_0, Spearman correlation, top-k recall, tail RMSE, and calibration error explains $R^2 = 0.411$ of the observed variation. Adding $R_{1,TT}$ and $R_{2,4\mathrm{held}}$ increases the in-sample explained variation to $R^2 = 0.856$, corresponding to

partial $R^2 = 0.756$. The within-system Freedman–Lane permutation test gives $p = 5.0 \times 10^{-4}$, while the system-aware bootstrap interval for partial R^2 is $[0.0046, 0.7564]$.

Because the analysis contains only two independent biological systems, we performed additional robustness analyses. Under leave-one-replicate-out evaluation using the original predictor blocks, adding both relational predictors reduces out-of-fold residual error by approximately 75.9% relative to the conventional reduced model.

This magnitude changes substantially after explicitly accounting for biological system and baseline-versus-intervention model role. Under these controls, the in-sample partial R^2 of the relational block falls to 0.058, while the leave-one-replicate-out residual reduction falls to approximately 2.75%. Considering the relational predictors separately under the same stricter controls gives out-of-fold residual reductions of approximately 3.07% for $R_{1,TT}$ and 2.51% for $R_{2,4\mathrm{held}}$.

The frozen primary analysis and the additional sensitivity analysis therefore support different levels of interpretation. Relational quantities are associated with downstream outcomes beyond the original conventional predictor block, but the very large partial R^2 from the predefined model is not robust to explicit system and model-role control. The result is therefore treated as a decision-relevant association within the evaluated setting rather than evidence of a large independent predictive effect, cross-system generalization, or causal improvement in sequential discovery.

The intervention analysis is separately negative under its predefined criterion. Mean system-level decision benefit is 0.0434, below the required 0.05, with one-sided 95% confidence interval $[-0.0692, 0.1560]$ and $p = 0.124$.

The sequential analysis was conditional on the evaluated acquisition policies. Global greedy acquisition is the predefined primary policy, with random acquisition retained as a comparator. Additional adaptive policies, including uncertainty-aware acquisition, local ascent, and Thompson sampling, were not evaluated.

C.10 Learned-direction exact matched-R_0 calibration and structural reconstruction

The deterministic constructions above establish an exact mechanism using predefined spectral residual directions. We therefore performed a separate secondary analysis using the residual directions actually produced by all 84 frozen primary-benchmark predictors. For model m, define

$$e_m(v) = \hat{f}_m(v) - f(v).$$

For target amplitude $\tau \in \{0.10, 0.20, 0.30\}$, we form

$$c_{m,\tau} = \frac{\tau s_G}{\mathrm{RMSE}_{\mathrm{test}}(e_m)}, \qquad \widetilde{f}_{m,\tau}(v) = f(v) + c_{m,\tau} e_m(v).$$

Because $c_{m,\tau} > 0$, this transformation preserves the learned residual direction and the componentwise residual sign pattern while removing between-model differences in residual magnitude. The 84 fixed residual directions and three amplitudes yield 252 calibrated reconstructions.

Held-out labels are used only to impose the exact normalized test-risk constraint and to evaluate the already declared held-out endpoints; they are not used to train, select, or tune any predictor. First-order evaluation uses $R_{1,TT}$ on test–test mutation edges. Second-order evaluation uses $R_{2,4\mathrm{held}}$ on mutation squares whose four corners are held out. Full-landscape GraphFLA descriptors are reported separately as reconstruction diagnostics rather than held-out topology-generalization endpoints. Seeds are within-system stability measurements, and amplitudes are experimental conditions rather than independent replicates.

Support and numerical integrity. The second-order audit uses all observed physical mutation squares rather than sampling: 84,456 on CPEB3, 70,713 on Domingo tRNA, 53,237 on F1*U, and 3,798,289 on Tetrahymena. The number of strictly four-held-out squares varies with the frozen primary split: 85–98 on CPEB3, 103–118 on Domingo tRNA, 71–83 on F1*U, and 6,066–6,079 on Tetrahymena. The known square-count gate is satisfied in every system and the calibrated R_0 values reproduce their targets to machine precision. Across the 252 GraphFLA reconstructions, the maximum absolute matched-R_0 error is 1.11×10^{-16}, with no input-row or retained-vertex support failures. Recomputing the original first-order quantities from the frozen primary-benchmark artifacts agrees with the stored $R_{1,TT}$ values to approximately 8.4×10^{-15}.

Table 13: Learned-direction relational separation at exact $R_0 = 0.20$. Ranges are over the three same-system, same-seed blocks. $R_{1,TT}$ uses held-out test–test mutation edges and $R_{2,4\text{held}}$ uses strictly four-held-out mutation squares.

System	max/min $R_{1,TT}$	max/min $R_{2,4\text{held}}$	$\lvert\mathbf{Q}_{4\text{held}}\rvert$
CPEB3	1.41–1.47×	2.58–2.82×	85–98
Domingo tRNA	1.27–1.28×	1.42–1.50×	103–118
F1*U	1.13–1.15×	1.38–1.42×	71–83
Tetrahymena	1.10–1.11×	1.15–1.17×	6066–6079

Table 14: Within-block spread in held-out mutation-sign accuracy at exact $R_0 = 0.20$. Values are percentage points across the twelve system–seed blocks.

γ	Mean spread (pp)	Maximum spread (pp)
0.10	3.328	7.269
0.25	0.813	2.013
0.50	0.095	0.387
1.00	0.000	0.000

Held-out relational separation under exact pointwise matching. At exact $R_0 = 0.20$, model classes retain different relational error within the same biological system and seed. Table 13 reports the range, over the three seed-matched blocks, of the within-block maximum to minimum ratio across the seven fixed model classes.

The second-order separation is strongest on CPEB3, where the three same-system, same-seed ratios are 2.583, 2.637, and 2.817. The corresponding ranges are 1.424–1.495 on Domingo tRNA, 1.377–1.419 on F1*U, and 1.147–1.169 on Tetrahymena. These are counterfactual magnitude-normalized comparisons: they show that differences in learned residual direction remain visible after pointwise error is forced to be exactly equal, not that independently trained models naturally attain identical R_0.

Margin-dependent mutation-sign recovery. We next restrict held-out test–test mutation edges to the true-effect margin set $M_{1,\gamma}$ defined in Appendix A.1. At exact $R_0 = 0.20$, the within-system, within-seed spread in mutation-sign accuracy across the seven learned residual directions decreases sharply as the true margin increases.

Table 14 summarizes this margin-dependent spread in held-out mutation-sign accuracy.

The attenuation is not explained simply by disappearance of the evaluated support. Table 15 reports the mean fraction of test–test edges retained after applying each true-effect margin. Even at $\gamma = 1.0$, approximately one quarter to one third of the held-out edge support remains, depending on system.

As a direct implementation audit of the margin bound, we evaluated all four numerical margins across four systems, seven models, three seeds, and three amplitudes, for 1,008 model–margin rows. The bound derived in Appendix A.1 holds in all 1,008 cases. This numerical check is not an independent statistical test of the theorem; it verifies that the implemented support and normalization conventions agree with the stated bound.

Reconstructed landscape structure. The 252 calibrated prediction fields are also evaluated with the frozen GraphFLA pipeline. Table 16 summarizes within-system, within-seed structural spread across model classes as a function of exact pointwise-error amplitude. Local-optima summaries use all 12 system–seed blocks. Accessibility is undefined for the fragmented F1*U landscape and therefore uses the remaining nine blocks.

At $R_0 = 0.20$, the mean local-optima-ratio spread is 0.01166 and the mean spread in absolute local-optimum distortion is 0.00964. Across the nine accessibility-eligible blocks, the corresponding mean raw accessibility spread is 0.13307 and the mean spread in absolute accessibility distortion is 0.09606. The descriptive increase from $R_0 = 0.10$ to 0.30 is consistent with larger residual amplitude amplifying structural consequences, while the exact pattern remains descriptor- and landscape-dependent.

Table 15: Mean retained test–test mutation-edge fraction by true-effect margin at exact $R_0 = 0.20$. The margin support is determined by the true landscape and frozen split, so it is common across model classes within a system–seed block.

System	$\gamma = .10$	$\gamma = .25$	$\gamma = .50$	$\gamma = 1.0$
CPEB3	76.75%	55.64%	39.62%	24.64%
Domingo tRNA	89.65%	74.79%	54.69%	26.44%
F1*U	49.73%	40.95%	32.90%	24.15%
Tetrahymena	91.42%	78.82%	60.06%	31.99%

Table 16: Amplitude dependence of learned-direction structural reconstruction. Entries are mean within-block spreads across model classes. "LO distortion" and "accessibility distortion" are absolute deviations from the corresponding measured-landscape descriptor.

R_0	LO ratio	LO distortion	Accessibility	Access. distortion
0.10	0.00821	0.00732	0.07534	0.05910
0.20	0.01166	0.00964	0.13307	0.09606
0.30	0.01453	0.01134	0.15281	0.11448

Together, the learned-direction analysis bridges the deterministic mechanism construction and naturally trained predictors. It shows across four RNA landscapes that residual directions learned by different model classes can retain different held-out relational errors, margin-dependent mutation-sign recovery, and reconstructed topology even after pointwise residual magnitude is forced to be identical. The analysis remains secondary and non-confirmatory: it does not estimate the prevalence of naturally matched-risk model pairs and does not convert the 252 calibrated reconstructions into independent biological replications.

D Scope and limitations of the combined empirical evidence

The experiments distinguish pointwise prediction error, relational-error magnitude, residual-frequency organization, and reconstructed combinatorial structure. The primary matched-risk benchmark does not establish the prespecified large first-order separation at naturally matched pointwise risk, and its observed $R_{1,TT}$ separations are not assigned standalone biological significance.

The mutation-square analysis provides stronger within-landscape structural evidence: nearly identical aggregate R_2 coexists with different interaction-sign and epistatic-class recovery and with markedly different graph-Fourier residual-frequency organization. Post-freeze paired inference shows that the principal sign-based structural contrasts are directionally consistent across the repeated seed–holdout configurations, while the conservative five-seed exact analysis is limited by the coarse attainable p-value resolution. The spectral extension is exploratory and uses the same fixed 30 prediction fields; its paired direction counts are not treated as independent biological replication or as formal inferential confirmation. The accompanying GraphFLA analysis shows that these differences can coincide with different reconstructed local-optimum and accessibility structure, although not every landscape descriptor improves in the same direction.

The controlled matched-R_0 constructions address a distinct mechanism question by removing aggregate pointwise magnitude as a confound. Exactly equal pointwise error produces substantially different first- and second-order relational errors and different reconstructed topology on Domingo tRNA, and the same separation mechanism is reproduced under the frozen extension protocol on Tetrahymena. These are deterministic mechanism demonstrations on two empirical RNA landscapes. They establish existence and cross-landscape portability of the construction, not prevalence or independent confirmatory biological replication; the three amplitudes within each landscape are conditions rather than replicates.

The learned-direction matched-R_0 analysis strengthens that bridge without changing the inferential status of the paper. It uses all 84 frozen primary-benchmark residual directions across four RNA landscapes and rescales each direction to three exact pointwise-error amplitudes. At matched magnitude, learned residual directions retain different held-out $R_{1,TT}$ and $R_{2,4\mathrm{held}}$, with the largest second-order separation on CPEB3, and mutation-sign differences concentrate near smaller true-effect margins. The full-landscape GraphFLA reconstructions also differ and become more sepa-

rated progressively as amplitude increases. These 252 reconstructions are counterfactual conditions, not independently trained equal-risk models or 252 biological replications. Full-landscape topology remains a reconstruction diagnostic rather than a held-out topology-generalization endpoint.

Relation-aware training provides complementary intervention evidence: joint relation-aware training reduces relational risk in both tested biological systems, although formal pointwise equivalence is not established. The subsequent objective decomposition further shows that first- and second-order preservation are not interchangeable: second-order-only preservation lowers R_2 in both systems but produces system-dependent effects on R_1 and the combined relational objective. The weighting sensitivity shows that the joint-preservation contrast is not an artifact of the equal-weight scalarization over the broad range examined, while also reinforcing that R_1 and R_2 should remain separately reported. The accompanying Pareto audit shows heterogeneous pointwise–relational trade-offs across validation-eligible selected objectives and demonstrates that relational improvement does not require pointwise degradation in every split–seed unit. Neither the decomposition, weighting sensitivity, nor the Pareto audit redefines the original frozen E5 confirmatory analysis or the failed system-level equivalence criterion.

The sequential-decision analysis shows an association between relational metrics and decision outcomes beyond the original conventional predictor block, but its magnitude is substantially attenuated after explicit control for biological system and model role, and the predefined decision-improvement criterion remains unestablished.

These findings do not imply that relational metrics universally dominate pointwise metrics. Pointwise accuracy remains essential, relational RMSE magnitude alone is not treated as evidence of structural importance, and the decision analysis is conditional on the evaluated acquisition policy. The evidence is also restricted to the declared combinatorial neighborhood structures and biological systems studied, which are predominantly RNA landscapes, with GB1 providing narrower cross-domain protein evidence.