VALIDATING INTERPRETABILITY IN siRNA EFFICACY PREDICTION: A PERTURBATION-BASED, DATASET-AWARE PROTOCOL

Zahra Khodagholi*
Department of Industrial Engineering
University of Central Florida
Zahra.Khodagholi@ucf.edu

Niloofar Yousefi
Department of Industrial Engineering
University of Central Florida
Niloofar.Yousefi@ucf.edu

ABSTRACT

Saliency maps are increasingly used as *design guidance* in siRNA efficacy prediction, yet attribution methods are rarely validated before motivating sequence edits. We introduce a **pre-synthesis gate**: a protocol for *counterfactual sensitivity faithfulness* that tests whether mutating high-saliency positions changes model output more than composition-matched controls. Cross-dataset transfer reveals two failure modes that would otherwise go undetected: *faithful-but-wrong* (saliency valid, predictions fail) and *inverted saliency* (top-saliency edits less impactful than random). Strikingly, models trained on mRNA-level assays collapse on a luciferase reporter dataset, demonstrating that protocol shifts can silently invalidate deployment. Across four benchmarks, 19/20 fold instances pass; the single failure shows inverted saliency. A biology-informed regularizer (BioPrior) strengthens saliency faithfulness with modest, dataset-dependent predictive trade-offs. Our results establish saliency validation as essential pre-deployment practice for explanation-guided therapeutic design. Code is available at https://github.com/shadi97kh/BioPrior.

1 INTRODUCTION

Small interfering RNAs (siRNAs) enable programmable, sequence-specific gene silencing and have become a practical modality in both therapeutic development and functional genomics (Elbashir et al., 2001; Fire et al., 1998). The clinical success of FDA-approved siRNA drugs, including patisiran, givosiran, and inclisiran (Adams et al., 2018; Balwani et al., 2020; Ray et al., 2020; Setten et al., 2019), has intensified interest in computational methods for predicting siRNA efficacy. In discovery settings, researchers routinely screen many candidate oligonucleotides and prioritize those expected to achieve strong knockdown. This has driven sustained interest in machine learning models that predict siRNA efficacy directly from nucleotide sequence and related descriptors (Han et al., 2018; Bai et al., 2024). In practice, this means: pick sequences, edit motifs or seed composition, adjust GC balance, and re-screen, so explanation quality directly affects experimental cost and iteration speed. Reliable saliency maps would enable practitioners to rationally edit candidate siRNA sequences at positions most likely to improve knockdown, accelerating the design of effective therapeutic oligonucleotides while reducing costly experimental iterations.

Modern deep predictors can be accurate on standard benchmarks, but an increasingly important question is whether they are *trustworthy* as decision-support tools. In practice, investigators do not only use predicted efficacy scores; they often inspect saliency maps or other attribution visualizations to infer which nucleotides "matter," and then use those attributions to motivate sequence edits (e.g., adjusting seed composition, GC balance, or motif avoidance). If these explanations are not faithful (meaning interventions at highlighted positions do not produce larger prediction changes than controls), then explanation-guided design can be misleading (Adebayo et al., 2018; Kindermans et al., 2019), especially under the protocol and distribution shifts that are common across assays, labs, and readouts.

*Corresponding author: Zahra.Khodagholi@ucf.edu

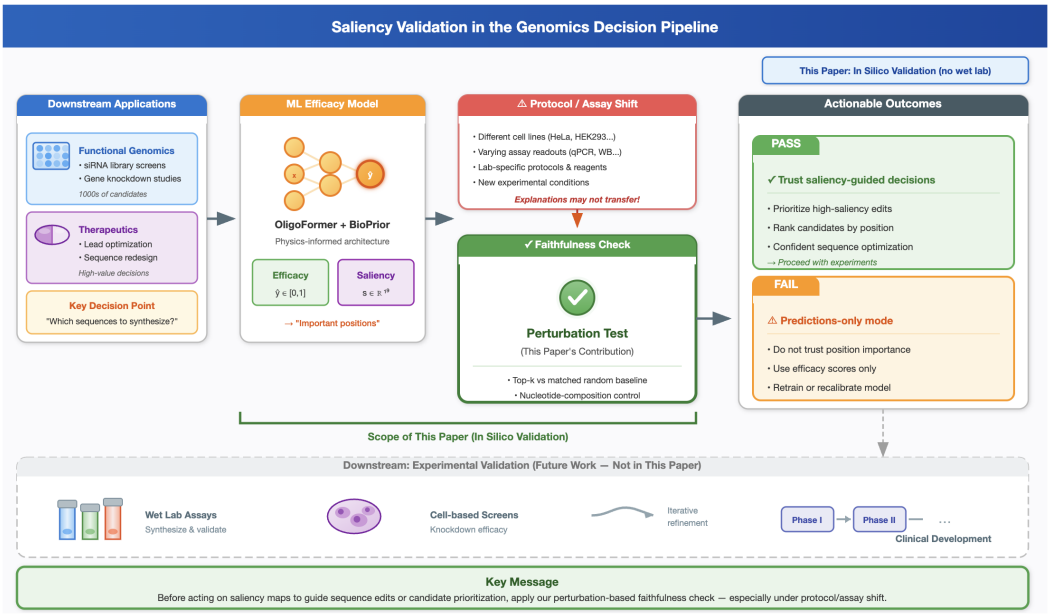


Figure 1: **Positioning saliency validation in the lab-in-the-loop decision pipeline.** In both therapeutic lead selection and functional genomics knockdown screens, researchers rely on predicted efficacy and position-level saliency ("important positions") to decide which siRNA sequences to synthesize or prioritize for experimental validation. However, explanation methods can appear plausible while failing basic perturbation tests, a risk that compounds under assay or protocol shift across laboratories, cell lines, and readout technologies. This paper introduces a standardized faithfulness check (expected-effect perturbations with a nucleotide-matched baseline) that practitioners can apply as a **pre-synthesis gate** before acting on saliency maps in a new dataset or experimental setting. When validation passes, saliency-guided decisions (sequence edits, candidate ranking) can be trusted; when it fails, predictions may still be useful but position-importance reasoning should be avoided. Downstream wet-lab validation and clinical development (dashed region) are outside the scope of this work.

We therefore focus on a concrete, testable desideratum: *a saliency method is useful for design only if mutating high-saliency positions changes the model's prediction more than mutating appropriate controls* (Samek et al., 2017). We term this *counterfactual faithfulness*: the saliency map correctly identifies positions where the model is sensitive to interventions. We propose this test as a **pre-synthesis gate**, a validation step practitioners should run before acting on saliency maps in explanation-guided siRNA design. Crucially, this is a **model-centric guarantee**: saliency tracks where the model is sensitive, not necessarily what biology truly cares about. This is distinct from *biological causality* (whether position changes affect true efficacy) and from *distributional faithfulness* (whether saliency reflects learned correlations). Our test validates model sensitivity under single-base perturbations, which is the operationally relevant property for explanation-guided sequence editing.

We introduce a perturbation-based validation protocol that operationalizes this idea for nucleotide sequence predictors. Given a trained model and a held-out siRNA, we (i) compute position-wise saliency on nucleotide identity channels (Simonyan et al., 2014; Sundararajan et al., 2017), (ii) select the top-k salient positions, (iii) quantify an expected-effect mutation score by averaging the prediction change under all single-base substitutions at those positions, and (iv) compare this score to a nucleotide-matched random baseline to control for compositional bias. This yields a simple pass/fail faithfulness test that can be run before saliency maps are used for design guidance (Amorim et al., 2023). Our protocol's key innovations over standard in-silico mutagenesis and faithfulness metrics are detailed in Section 2 and Table 1.

In parallel, we propose a biology-informed hybrid model for siRNA efficacy prediction that incorporates established design principles as differentiable regularizers. These priors include thermodynamic asymmetry (Schwarz et al., 2003), seed-region composition constraints (Jackson & Linsley, 2010),

global GC heuristics (Reynolds et al., 2004; Ui-Tei et al., 2004), immune-motif avoidance (Judge et al., 2005), and a duplex stability proxy penalizing excessive GC content that may indicate inaccessible targets. This approach aligns with recent work on physics-informed machine learning (Raissi et al., 2019; Karniadakis et al., 2021) and biologically interpretable neural networks (Elmarakeby et al., 2021; Novakovsky et al., 2023; Chen et al., 2024), though biological systems present unique challenges compared to physical systems due to uncertain priors and context-dependent mechanisms (Martinelli, 2025).

Across four benchmark datasets, we find that saliency faithfulness holds in 95% of fold–dataset settings and that salient positions cluster in canonical functional regions (seed and 3′ end). However, cross-dataset transfer reveals strong protocol dependence: models transfer among three datasets but fail on a luciferase-reporter dataset, suggesting assay-specific confounds that can break generalization even when within-dataset explanations are faithful. Our central thesis is that **saliency should be treated as a deploy-time claim**: validate explanation faithfulness on the target assay before using it to guide siRNA selection or modification.

Contributions.

1. **Introduce** a composition-controlled, perturbation-based protocol to validate saliency faithfulness for nucleotide sequence predictors, positioned as a pre-synthesis gate in lab-in-the-loop design workflows.
2. **Demonstrate** that validated saliency aligns with biologically meaningful siRNA regions across multiple benchmarks (19/20 fold–dataset combinations pass), with high-saliency positions clustering at the 5′ and 3′ termini adjacent to known functional determinants.
3. **Characterize** two distinct transfer failure modes: *faithful-but-wrong* (saliency valid, predictions fail) and *inverted saliency* (high-saliency positions less important than random), highlighting when explanations should not be trusted without dataset-specific validation.
4. **Show** that a biology-informed regularizer (BioPrior) strengthens explanation faithfulness with dataset-dependent predictive effects, demonstrating that mechanism-informed training can improve saliency reliability.

We release code and the validation protocol to enable adoption across sequence modeling applications.[1]

2 Related Work

siRNA efficacy prediction: from rules to deep learning. Early siRNA design relied on empirically derived rules capturing thermodynamic asymmetry, seed-region composition, and GC-related heuristics (Ui-Tei et al., 2004; Reynolds et al., 2004; Amarzguioui & Prydz, 2004). More recent approaches learn nonlinear sequence-to-efficacy mappings using deep neural networks, including transformer-based models that integrate thermodynamic descriptors and pretrained RNA sequence representations for improved accuracy and transfer (Han et al., 2018; Bai et al., 2024), graph neural networks that capture siRNA-mRNA interaction dynamics (Long et al., 2024), and preference-based ranking frameworks that address dataset bias through debiased training objectives (Zhang et al., 2025).

Biology-informed regularization for sequence models. In scientific ML, incorporating domain knowledge as soft constraints or regularizers can improve robustness and align learned behavior with known mechanisms (Raissi et al., 2019; Karniadakis et al., 2021). However, biological systems present unique challenges compared to physical systems: uncertain and context-dependent prior knowledge, heterogeneous and noisy data, partial observability, and complex high-dimensional networks (Martinelli, 2025). We adopt a biology-informed approach for siRNA prediction by encoding established design principles as differentiable penalties rather than hard constraints, allowing the model to learn dataset-specific deviations from canonical rules while remaining grounded in mechanistic understanding. This aligns with recent work on biologically interpretable neural networks

[1] Code available at `https://github.com/shadi97kh/BioPrior`.

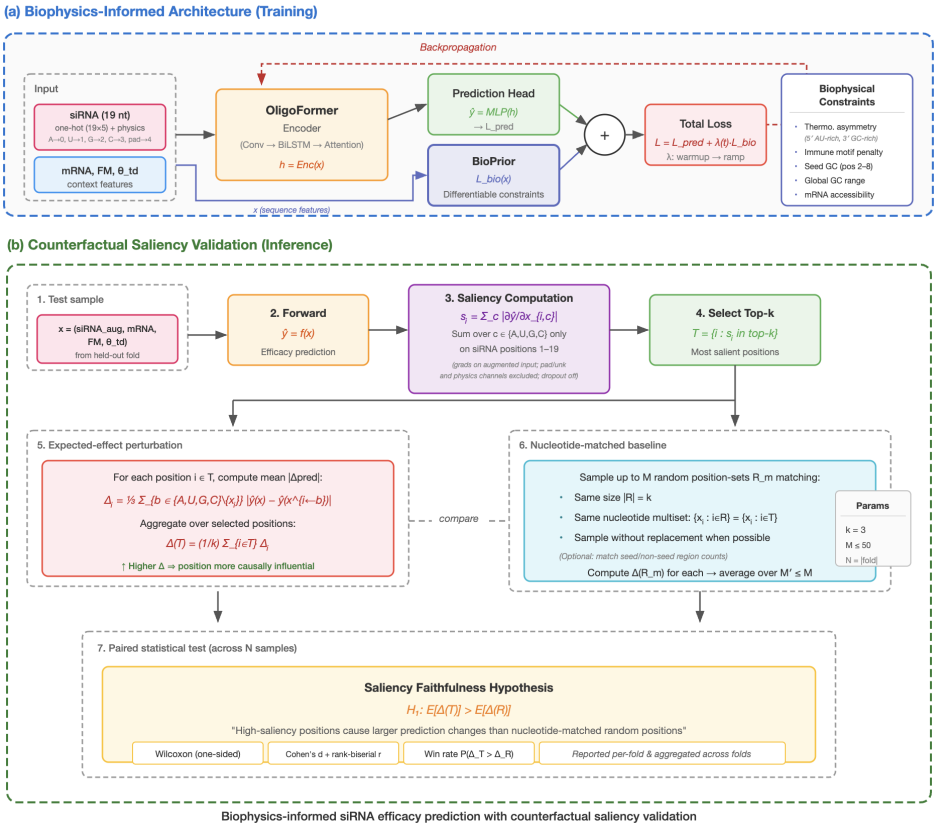


Figure 2: **Overview of training and saliency faithfulness. (a)** A hybrid Conv–BiLSTM–Transformer with dual siRNA↔mRNA cross-attention predicts efficacy from sequence encodings plus RNA-FM and thermodynamic features, regularized by BioPrior constraints weighted by a schedule $\lambda(t)$. **(b)** Saliency is validated by perturbing top-k salient siRNA positions (A/U/G/C channels only) and comparing the expected prediction change to a nucleotide-matched random baseline.

that maintain both predictive power and scientific transparency (Elmarakeby et al., 2021; Novakovsky et al., 2023; Chen et al., 2024).

Saliency methods and faithfulness validation. Gradient-based saliency methods attribute predictions to input features via derivatives and variants such as integrated gradients (Simonyan et al., 2014; Sundararajan et al., 2017). However, saliency maps are not guaranteed to reflect true feature importance (Adebayo et al., 2018; Kindermans et al., 2019; Rudin, 2019); a standard validation approach is perturbation-based testing, which measures whether modifying high-attribution features produces larger output changes than modifying suitable controls (Samek et al., 2017; Amorim et al., 2023).

Interpretability validation in sequence biology. In genomics, *in-silico mutagenesis* (ISM) interprets sequence models by mutating positions and measuring prediction changes (Alipanahi et al., 2015; Zhou et al., 2018). Related approaches include motif discovery from learned filters (Lanchantin et al., 2017), TF-MoDISco (Shrikumar et al., 2018), and deletion/insertion metrics (Hooker et al., 2019). Our protocol differs from both ISM and standard faithfulness evaluation (Samek et al., 2017; Hooker et al., 2019) in four ways: (1) an expected-effect operator averaging over all 3 substitutions per position; (2) a composition-matched baseline controlling for nucleotide-specific sensitivity; (3) explicit pass/fail acceptance criteria for deploy-time decisions; and (4) a cross-dataset diagnostic taxonomy (faithful-but-wrong vs. inverted-saliency). These components are individually known but their combination for nucleotide sequences is novel; Table 1 summarizes the ISM comparison.

3 BACKGROUND

Task and datasets. We study siRNA efficacy prediction from 19-mer guide sequences, where efficacy $y \in [0, 1]$ reflects target knockdown strength. We index siRNA positions 1–19 from 5′ to 3′ throughout this paper. We evaluate on four public benchmarks spanning distinct experimental protocols and distribution shift (Hu, Taka, Mix, Shabalina; Table 2), and normalize efficacy scores to $[0, 1]$. Following common practice, we define the high-efficacy regime as $y \geq 0.7$ and report both continuous (correlation) and thresholded (AUC) metrics.

Gradient-based saliency. Given a trained predictor $f_\theta(\mathbf{x})$ and an siRNA input representation with nucleotide channels $\mathbf{X}^{\text{si}} \in \mathbb{R}^{19\times C}$, we compute position-wise saliency using the gradient magnitude restricted to nucleotide identity (A/U/G/C) channels:

$$s_i = \sum_{c\in\{A,U,G,C\}} \left| \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{X}^{\text{si}}_{i,c}} \right|. \tag{1}$$

Faithfulness desideratum. A saliency map is *faithful* if perturbing high-saliency positions causes larger prediction changes than perturbing appropriate controls (Samek et al., 2017). We operationalize this via counterfactual single-nucleotide substitutions and compare the expected prediction change for top-k salient positions against a nucleotide-matched random baseline (Section 4.4).

Faithfulness taxonomy. To avoid confusion, we distinguish three notions:

Term	Definition	Validated by
Sensitivity faithfulness	High-saliency positions change model output more than matched controls	Perturbation test (this work)
Causal faithfulness	Interventions change ground-truth efficacy	Wet-lab experiments
Stability	Explanations invariant to irrelevant transforms	Input perturbation

Our test validates **sensitivity faithfulness**, the operationally relevant property for explanation-guided sequence editing. We also introduce terminology for transfer failures:

- **Faithful-but-wrong:** Saliency test passes but predictions fail (model is internally consistent but learned wrong rules).
- **Inverted saliency:** Saliency test fails with $d_z < 0$ (high-saliency positions are *less* important than random).

	ISM	Ours
Purpose	Interpretation	Validation
Output	Mutation effect profile	Pass/fail decision
Baseline	None (raw effects)	Composition-matched
Statistical test	Optional	Required (Wilcoxon)
Intended use	Post-hoc explanation	Deploy-time gate

Table 1: Comparison: In-silico mutagenesis (ISM) vs. our protocol.

Unlike ISM, which is an interpretability *output*, our protocol is a statistical *acceptance test* that decides whether explanations are safe to use for editing under a given dataset.

4 METHOD

We propose a biology-informed architecture for siRNA efficacy prediction that integrates established design principles as differentiable regularizers and introduces a counterfactual perturbation protocol to validate gradient-based saliency. Concretely, our method extends OligoFormer-style sequence modeling with (i) a hybrid encoder combining convolution, BiLSTM, and self-attention, (ii) dual-stream siRNA–mRNA cross-attention fusion, (iii) a BioPrior module that computes differentiable

constraint penalties from sequence-derived features (optionally conditioned on per-position nucleotide probabilities), and (iv) a saliency faithfulness test based on an expected-effect perturbation operator with a nucleotide-matched random baseline.

Overview. Figure 2 summarizes our approach. During training (Fig. 2a), we optimize a predictive model while softly enforcing established siRNA design principles via a differentiable BioPrior regularizer whose contribution is scheduled over epochs. During evaluation (Fig. 2b), we validate whether gradient-based saliency exhibits *counterfactual faithfulness* using a perturbation test that compares the effect of mutating top-saliency positions against a nucleotide-matched baseline.

4.1 Model Architecture

Overview. Our backbone follows **OligoFormer** (Bai et al., 2024): a hybrid **Conv** $\rightarrow$ **BiLSTM** $\rightarrow$ **attention** encoder for siRNA and mRNA, followed by **bidirectional cross-attention** and an MLP prediction head. The appendix provides full architectural hyperparameters and implementation details (Appendix H).

Inputs. Each example contains (i) a 19-nt siRNA guide, (ii) an mRNA context window centered at the target site, and (iii) global descriptors. We encode siRNA as $\mathbf{X}^{\text{si}} \in \mathbb{R}^{19\times C_{\text{si}}}$ where $C_{\text{si}} = 13$ includes a 5-channel one-hot (A,U,G,C,pad) plus lightweight sequence-derived indicator channels (e.g., seed/cleavage flags and AU/GC indicators). We encode mRNA as $\mathbf{X}^{\text{mr}} \in \mathbb{R}^{L_{\text{mr}}\times 5}$ (A,U,G,C,pad). We additionally use pooled RNA-FM embeddings $\mathbf{z}^{\text{si}}_{\text{FM}}, \mathbf{z}^{\text{mr}}_{\text{FM}}$ (Chen et al., 2022) and a thermodynamic descriptor vector $\mathbf{z}_{\text{td}} \in \mathbb{R}^{d_{\text{td}}}$ computed from sequence and thermodynamic tools (Lorenz et al., 2011).

Encoders and fusion. Both siRNA and mRNA are encoded with a shallow convolutional front-end, a 2-layer BiLSTM (Hochreiter & Schmidhuber, 1997), and a lightweight self-attention block (Vaswani et al., 2017), producing contextual representations $\tilde{\mathbf{H}}^{\text{si}}$ and $\tilde{\mathbf{H}}^{\text{mr}}$. We then fuse the two streams using **dual cross-attention** (siRNA$\leftarrow$mRNA and mRNA$\leftarrow$siRNA), pool each stream by mean+max pooling, concatenate global descriptors $(\mathbf{z}^{\text{si}}_{\text{FM}}, \mathbf{z}^{\text{mr}}_{\text{FM}}, \mathbf{z}_{\text{td}})$, and map the result to the final efficacy prediction via an MLP.

Per-position nucleotide probability head. To couple the predictor to differentiable mechanistic constraints, we attach a small auxiliary head that outputs per-position nucleotide probabilities $\mathbf{P}^{\text{si}} \in [0,1]^{19\times 4}$ from intermediate siRNA features (after the BiLSTM encoder). These probabilities are computed via softmax over learned logits $\mathbf{z} \in \mathbb{R}^{19\times 4}$:

$$P_{i,b} = \frac{\exp(z_{i,b})}{\sum_{b'\in\{A,U,G,C\}} \exp(z_{i,b'})}$$

The probabilities are used only by BioPrior (below) and do not alter the main prediction pathway.

4.2 Biology-Informed Regularization (BioPrior)

We encode established siRNA design principles as differentiable penalties computed from deterministic sequence-derived features and optional soft nucleotide probabilities $\mathbf{P}^{\text{si}}$. The BioPrior module aggregates constraint losses:

$$\mathcal{L}_{\text{bio}} = \sum_{c\in\mathcal{C}} \bar{\alpha}_c \, \mathcal{L}_c, \tag{2}$$

where $\mathcal{C}$ includes (i) terminal asymmetry preference, (ii) seed composition constraints, (iii) global GC constraints, (iv) immune motif avoidance, and (v) a duplex stability proxy based on siRNA GC content. We compute BioPrior on explicit sequence-level quantities (rather than latent representations) to avoid injecting spurious position biases. A full mathematical specification and proofs are provided in Appendix B.

Scheduling. We introduce BioPrior with an epoch-based warmup-and-ramp:

$$\lambda(t) = \begin{cases} 0 & t < t_{\text{warm}}, \\ \min(\lambda_{\max}, \lambda_0 + \gamma(t - t_{\text{warm}})) & \text{otherwise}, \end{cases} \tag{3}$$

with $t_{\text{warm}} = 8$ epochs, $\lambda_0 = 0.10$, $\gamma = 0.01$ per epoch, and $\lambda_{\max} = 0.30$. This schedule begins at $\lambda_0 = 0.10$ after warmup and ramps linearly, allowing the model to first learn predictive features before gradually increasing biological regularization. The relatively high cap accommodates the normalized constraint weights (Section 4.2), which distribute the effective penalty across five terms; the per-constraint contribution remains moderate throughout training.

4.3 Training Objective

The overall objective combines prediction, BioPrior regularization, and auxiliary mechanistic supervision:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda(t)\,\mathcal{L}_{\text{bio}} + \lambda_{\text{aux}}\,\mathcal{L}_{\text{aux}}. \tag{4}$$

Gradient flow through soft probabilities. The model produces soft nucleotide probability distributions $P \in \mathbb{R}^{L\times 4}$ (where $P_{i,b} = \Pr(\text{position } i = b)$, $\sum_b P_{i,b} = 1$) alongside efficacy predictions. **In all main experiments**, BioPrior penalties are computed on these soft probabilities: $\mathcal{L}_{\text{bio}} = \mathcal{L}_{\text{bio}}(P_{\text{model}})$. This ensures gradients from biological constraints backpropagate through the model's sequence representation. At inference with fixed sequences, one-hot encodings are used (gradient flow not needed). An ablation using deterministic one-hot BioPrior is in Appendix I.

We optimize with AdamW and standard regularization/early stopping under 5-fold cross-validation; the complete optimization protocol, hyperparameters, and compute details are listed in Appendix H.

4.4 Counterfactual Saliency Validation

We test whether gradient saliency highlights positions with high interventional sensitivity using a counterfactual perturbation protocol with a nucleotide-matched baseline. Unless stated otherwise, we compute saliency using the gradient magnitude restricted to the siRNA nucleotide identity channels (A/U/G/C), aligning attribution with the intervention space used in our perturbation test; Integrated Gradients is reported only as a robustness check (Appendix E). Specifically, saliency is computed with respect to the one-hot nucleotide channels at the model input layer; gradients flow through any learned projections in the standard manner. Derived channels (GC content, seed indicators, thermodynamic features) are excluded from attribution because they are deterministically recomputed after perturbation, so attributing to nucleotide identity captures the causal parent in the input graph.

Important: Gradient-based saliency measures model sensitivity to nucleotide channels *holding derived features fixed*; it is an approximation. The perturbation validation (Section 4.4) is our primary interpretability test because it recomputes all derived features after each mutation, capturing the true causal effect of nucleotide changes.

Given a held-out example, we compute position saliency s_i via Eq. 1, select the top-k positions T, and quantify their *expected effect* under all single-base substitutions:

$$\Delta_i = \frac{1}{3} \sum_{b\in\{A,U,G,C\}\setminus\{x_i\}} \left|\hat{y}(\mathbf{X}) - \hat{y}(\mathbf{X}^{i\leftarrow b})\right|, \qquad \Delta(T) = \frac{1}{k}\sum_{i\in T}\Delta_i. \tag{5}$$

After each substitution we recompute all derived input channels (seed indicators, GC content, thermodynamic asymmetry) to ensure input coherence; this deterministic recomputation can introduce nonlinear effects that break simple gradient-effect correspondence.

What is recomputed per mutation: Derived indicator channels and thermodynamic descriptors are recomputed after each substitution. RNA-FM embeddings are held fixed for tractability; we verify on a 50-sequence subset that recomputing FM changes d_z by < 0.05 (Appendix E). We test faithfulness under single-base edits, the operative action in sequence refinement, rather than distribution-preserving perturbations.

Averaging structure. Our test averages over three levels: (i) 3 substitutions per position (all bases except the original), (ii) k top-saliency positions, and (iii) M' composition-matched random sets. This yields one paired difference $d = \Delta(T) - \Delta_{\text{match}}$ per sequence, enabling standard paired statistical testing across the held-out set.

Algorithm: Counterfactual Faithfulness Test

Input: Model f_θ, held-out set $\mathcal{D}$, top-k, matched samples M'
Output: Pass/fail, effect size d_z, win rate

`for` each $(\mathbf{x}, y) \in \mathcal{D}$:

1. Compute saliency: $s_i = \sum_c |\frac{\partial \hat{y}}{\partial x_{i,c}}|$
2. Select top-k positions: $T = \text{argtop}_k(s)$
3. Compute expected effect: $\Delta(T) = \frac{1}{k} \sum_{i \in T} \frac{1}{3} \sum_{b \neq x_i} |\hat{y}(\mathbf{x}) - \hat{y}(\mathbf{x}^{i \leftarrow b})|$
4. Sample M' composition-matched random sets $\{R_m\}$
5. Compute baseline: $\Delta_{\text{match}} = \frac{1}{M'} \sum_m \Delta(R_m)$
6. Record difference: $d = \Delta(T) - \Delta_{\text{match}}$

`end for`

Compute Wilcoxon signed-rank p, Cohen's d_z, win rate
Pass if: $p < 0.05$ **and** $d_z > 0.2$ **and** win rate $> 50\%$

Why this test is nontrivial. Gradient-based saliency and perturbation effects are not guaranteed to align:

- **Discrete perturbations:** We perturb nucleotide identities, not continuous features; local linearity assumptions may fail.
- **Derived channel recomputation:** After each mutation we recompute seed indicators, GC content, and thermodynamic features, producing nonlinear, discontinuous changes that gradients do not capture.
- **Negative controls fail:** Randomized-weight, shuffled-label, shuffled-saliency, and bottom-k controls all fail the test (Table 6).
- **Transfer produces inverted saliency:** Taka→Hu yields $d_z = -1.25$, where high-saliency positions are *less* sensitive than random. This smoking-gun failure confirms the test has discriminative power.

Additional details on gradient saturation and intervention-space alignment are in Appendix E.

To control composition bias, we sample random position sets with the *same nucleotide multiset* as T and compute the matched baseline $\Delta(R)$. We additionally run a stricter *region+composition matched* baseline; faithfulness remains significant under this control (Table 12 in Appendix F). We evaluate faithfulness across held-out samples using paired one-sided Wilcoxon signed-rank tests (Wilcoxon, 1945) and report effect sizes.

Limitation: position vs. composition matching. Our primary baseline matches nucleotide composition but not positional region. If high-saliency positions cluster at inherently sensitive regions (e.g., sequence ends), this could inflate results. To address this concern, we conducted additional experiments with a *region-matched* baseline that samples random positions from the same coarse region buckets as top-k (5′ terminus: 1–4, seed-adjacent: 5–8, cleavage: 9–11, mid: 12–15, 3′ terminus: 16–19) while also matching nucleotide composition. Results remain significant: Hu achieves 78.3% win rate ($d_z = 0.71$) under region+composition matching versus 85.2% ($d_z = 0.86$) under composition-only matching. The reduced but still significant effect confirms that saliency captures position-specific patterns beyond regional sensitivity.

Additionally, our transfer analysis provides indirect evidence against positional bias: Taka-trained models show high saliency at positions 9–11, yet these positions show *inverted* sensitivity when tested on other datasets. If positional bias were dominant, we would expect consistent sensitivity regardless of training source. The asymmetric transfer failures suggest saliency captures learned position-specific patterns rather than inherent regional sensitivity.

The full mathematical specification is provided in Appendix C.

Faithfulness Test Specification

Test: Paired one-sided Wilcoxon signed-rank; H_1: $\mathbb{E}[\Delta(T)] > \mathbb{E}[\Delta_{\text{match}}]$
Defaults: $k = 3$, $N \geq 100$ samples, $M' = 50$ matched sets per sample
Pass criteria: $p < 0.05$ **and** $d_z > 0.2$ (Cohen, 1988) **and** win rate $> 50\%$
Threshold calibration: Trained models achieve $d_z = 0.70$–1.07; negative controls yield $d_z \in [-0.45, 0.03]$. The $d_z > 0.2$ threshold cleanly separates learned from spurious saliency.
Scope: Validates model sensitivity (where edits change predictions), **not** biological causality or OOD generalization.
If fail: Do not use saliency for design; check for distribution shift; retrain on protocol-matched data.

5 EXPERIMENTS

We evaluate our biology-informed sequence model on four benchmark siRNA datasets, assessing (i) predictive performance and (ii) explanation faithfulness using our counterfactual saliency validation protocol.

5.1 DATASETS

We use four publicly available siRNA efficacy datasets spanning different experimental protocols and assays (Table 2).

Huesken (Hu). The largest benchmark dataset, containing 2,431 siRNAs targeting 34 human genes measured in H1299 cells (Huesken et al., 2005). Efficacy was quantified via branched DNA assay measuring residual mRNA levels 24 hours post-transfection. Scores are normalized to $[0, 1]$ where higher values indicate stronger knockdown.

Katoh (Taka). A dataset of 702 siRNAs all targeting a single luciferase reporter mRNA, evaluated using dual-luciferase reporter assays in HeLa cells (Katoh & Suzuki, 2007). This dataset exhibits different distributional properties and position-importance patterns compared to Hu, likely reflecting differences in assay design and the single-target nature of the experiments.

Mix. Following Bai et al. (2024), we adopt the dataset categorization from OligoFormer, which collected nine datasets comprising 3,714 siRNAs and 75 mRNAs from prior studies, then organized them into three groups: the Huesken dataset, the Takayuki dataset, and the remaining seven studies merged into a single "Mixset." The resulting Mix dataset contains 581 siRNAs compiled from these seven independent studies with heterogeneous experimental protocols: Amarzguioui (46 siRNAs targeting 4 mRNAs in HaCaT cells) (Amarzguioui et al., 2003), Harborth (44 siRNAs targeting 1 mRNA in HeLa) (Harboth et al., 2003), Hsieh (108 siRNAs targeting 22 mRNAs in HEK293T) (Hsieh et al., 2004), Khvorova (14 siRNAs targeting 1 mRNA in HEK293) (Khvorova et al., 2003), Reynolds (240 siRNAs targeting 7 mRNAs in HEK293) (Reynolds et al., 2004), Vickers (76 siRNAs targeting 2 mRNAs in T24) (Vickers et al., 2003), and Ui-Tei (53 siRNAs targeting 3 mRNAs in HeLa) (Ui-Tei et al., 2004). Inhibition efficacy across all constituent datasets was normalized to $[0, 1]$, with 0.7 used as the threshold for high-efficacy classification, consistent with the OligoFormer preprocessing. Needleman–Wunsch global alignment (Needleman & Wunsch, 1970) was applied to remove redundant sequences (identity $> 80\%$) across splits, yielding the final dataset. While the aggregation provides scale and diversity across cell lines and target genes, the mixed provenance introduces additional noise and potential batch effects.

Shabalina. A curated dataset of 653 siRNAs aggregated from published literature with computational filtering for thermodynamic properties (Shabalina et al., 2006). The curation process may introduce selection biases favoring well-characterized sequences.

Threshold sensitivity. Because datasets are independently normalized to $[0, 1]$, the 0.7 threshold may correspond to different absolute knockdown levels across assays. We verified that transfer conclusions are robust to threshold choice: AUC rankings remain consistent at 0.6, 0.7, and 0.8 thresholds, and using top-30% per dataset as positives yields the same transfer patterns (Taka remains the outlier). Full sensitivity analysis is in Appendix I.

Dataset	N	Targets	Sources	Mean	Std	High-eff.	Assay
Hu	2,431	34	1	0.58	0.23	847 (34.8%)	bDNA
Taka	702	1	1	0.61	0.25	298 (42.5%)	Luciferase
Mix	581	40	7	0.56	0.24	183 (31.5%)	Mixed
Shabalina	653	>30	1	0.54	0.26	199 (30.5%)	Literature

Table 2: Dataset characteristics. High-efficacy threshold is $y \geq 0.7$.

5.2 Baselines and Ablations

OligoFormer baseline. We compare to the published OligoFormer model (Bai et al., 2024), which combines sequence representations with thermodynamic and foundation-model features.

Rule-based baseline. We include a classical thermodynamic scoring baseline based on established heuristic rules (e.g., Reynolds (Reynolds et al., 2004), Ui-Tei (Ui-Tei et al., 2004)).

Ablations. To isolate contributions of our components, we evaluate:

- **Baseline (no BioPrior):** identical architecture trained with prediction loss only (biology regularization disabled; $\lambda(t) = 0$).
- **+BioPrior (scheduled):** full model with biology loss $\mathcal{L}_{\text{bio}}$ enabled and scheduled via warmup–ramp $\lambda(t)$.

When reporting saliency faithfulness, we use the same faithfulness protocol for all variants to ensure comparability.

5.3 Evaluation Metrics

Predictive performance. We report:

- **MSE**: mean squared error on $y \in [0, 1]$.
- **Pearson** r and **Spearman** ρ: correlation between predicted and true efficacy.
- **ROC-AUC** and **PR-AUC**: treating $y \geq 0.7$ as the positive class.

Faithfulness. Using the protocol in Section 4.4, we report:

- **Win rate**: fraction of samples where $\Delta(T) > \Delta_{\text{match}}$ (top-k exceeds the nucleotide-matched baseline).
- **Paired effect sizes**: Cohen's d_z (paired standardized mean difference: $d_z = \bar{d}/s_d$) and rank-biserial correlation on paired differences.
- **Statistical significance**: one-sided paired Wilcoxon signed-rank test for $H_1 : \mathbb{E}[\Delta(T)] > \mathbb{E}[\Delta_{\text{match}}]$.

All metrics are reported as mean $\pm$ standard deviation across folds.

6 Results

We present results on predictive performance, cross-dataset generalization, saliency faithfulness validation, and ablation studies. All experiments use 5-fold cross-validation; we report mean $\pm$ std across folds. Transfer experiments were repeated with 3 random seeds per configuration; we report aggregated statistics with complete tables in Appendix I.

6.1 Intra-Dataset Predictive Performance

Table 3 summarizes 5-fold cross-validation performance. Our biology-informed model (+BioPrior) achieves consistent improvements over the baseline: +0.01 AUC, +0.02 PR-AUC, and +0.01 PCC on

Dataset	Model	AUC	PR-AUC	PCC	F1
Hu	Baseline	0.82±.01	0.81±.02	0.63±.02	0.77±.01
	+BioPrior	**0.83**±.02	**0.82**±.02	**0.64**±.02	0.77±.03
Mix	Baseline	0.80±.04	0.80±.06	0.60±.04	0.76±.06
	+BioPrior	**0.81**±.05	**0.81**±.07	**0.61**±.08	**0.77**±.06
Taka	Baseline	0.82±.05	0.63±.10	0.68±.08	0.59±.07
	+BioPrior	**0.84**±.05	**0.67**±.10	0.68±.08	**0.60**±.07
Shabalina	Baseline	0.72±.04	0.66±.07	0.49±.05	0.65±.06
	+BioPrior	0.72±.02	0.66±.06	0.49±.03	**0.68**±.03

Table 3: Intra-dataset predictive performance (5-fold CV). Best results in **bold**. PR-AUC included given class imbalance (30–42% positives).

Source	Target	Baseline	+BioPrior	Diff
Strong generalization (AUC $>$ 0.75)				
Shabalina	Mix	0.813	**0.816**	+0.3%
Mix	Hu	0.792	0.792	0.0%
Shabalina	Hu	0.786	**0.787**	+0.1%
Hu	Mix	0.775	0.773	-0.2%
Moderate generalization (AUC 0.65–0.75)				
Mix	Shabalina	0.713	0.712	-0.1%
Hu	Shabalina	0.699	0.698	-0.1%
Poor generalization (AUC $<$ 0.60)				
Shabalina	Taka	0.574	0.559	-1.5%
Hu	Taka	0.536	0.535	-0.1%
Mix	Taka	0.499	0.497	-0.2%
Failed transfer from Taka (inverted predictions)				
Taka	Shabalina	0.517	0.517	0.0%
Taka	Mix	0.510	0.510	0.0%
Taka	Hu	0.490	0.490	0.0%

Table 4: Inter-dataset transfer performance (AUC). Models trained on source dataset, evaluated on target.

average across datasets. At the fold level, BioPrior improves AUC in 15/20 fold-dataset combinations (75%), with the largest gains on Taka (+0.02 AUC, +0.04 PR-AUC). BioPrior also strengthens saliency faithfulness (Section 6.4).

6.2 Cross-Dataset Generalization

A critical test of biological validity is whether models generalize across datasets collected under different experimental conditions. Table 4 presents inter-dataset transfer results, revealing substantial heterogeneity in generalization patterns.

Key findings. Three patterns emerge from the transfer experiments:

Cross-dataset sequence overlap. Before interpreting transfer results, we verified that sequence overlap does not inflate performance. Using Needleman-Wunsch alignment with 80% identity threshold, we found minimal overlap: Hu-Mix share 12 sequences (0.5%), Hu-Shabalina share 8 (0.3%), and Taka shares $<$5 sequences with any other dataset. Transfer performance thus reflects genuine generalization rather than memorization of shared sequences.

(1) Asymmetric generalization. Transfer success is highly asymmetric. Shabalina→Mix achieves 0.816 AUC while Mix→Shabalina reaches only 0.713. This suggests dataset-specific biases that benefit transfer in one direction but not the reverse.

Dataset	Pass	Win %	d_z	p-value	Status
Hu	5/5	85.2$\pm$4.1	0.86$\pm$.26	$<$0.001	✓
Mix	5/5	83.7$\pm$6.2	0.93$\pm$.45	$<$0.001	✓
Taka	5/5	87.1$\pm$3.8	1.07$\pm$.24	$<$0.001	✓
Shabalina	4/5	81.4$\pm$12.3	0.70$\pm$.42	(per-fold)†	✓†

†4 folds: $p < 0.001$; fold 1 failed: $d_z = -1.20$, win rate 18.3%, $p = 0.99$.

Table 5: Intra-dataset saliency faithfulness (+BioPrior model, 5-fold CV).

Control	Win %	d_z	Pass?
Trained model (reference)	85.2	0.86	✓
Randomized weights	51.2	0.03	×
Shuffled labels	48.7	−0.05	×
Shuffled saliency	49.3	0.01	×
Bottom-k selection	32.1	−0.45	×

Table 6: Negative control validation (Hu dataset). All controls fail to meet pass criteria ($d_z > 0.2$, win rate $> 50\%$), confirming the test distinguishes learned from spurious saliency.

(2) Taka as systematic outlier. Models trained on Taka fail to generalize to any other dataset (AUC ≈ 0.50), and models trained on other datasets fail on Taka. The negative PCC values when Taka-trained models are applied elsewhere suggest inverted label relationships: sequences that Taka labels as high-efficacy may violate rules that predict efficacy in other datasets. Strikingly, Taka-trained models achieve their best transfer performance at epoch 0 (the untrained model), with test AUC *decreasing* during training (Figure 5a).

(3) Marginal effect of biology constraints on transfer. The biology-informed model shows small improvements on Shabalina-sourced transfers (+0.1–0.3% AUC) but minimal effect elsewhere, suggesting that mechanism-based constraints cannot rescue fundamentally misaligned datasets. A per-metric comparison of Baseline vs. +BioPrior across Mix-sourced transfers is in Appendix I.2 (Figure 9).

6.3 Saliency Faithfulness Validation

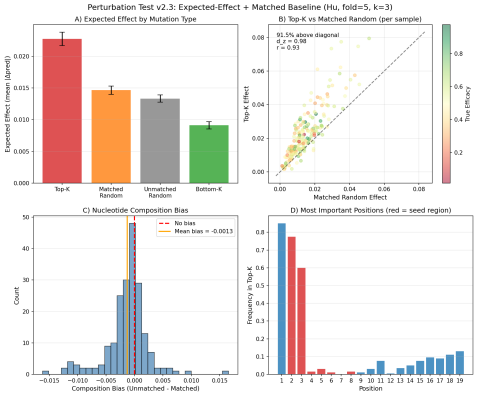
Our perturbation testing protocol validates whether gradient-based saliency maps identify positions with high interventional sensitivity. We evaluate faithfulness both within datasets (intra-dataset) and across datasets (inter-dataset transfer).

6.3.1 Intra-Dataset Faithfulness

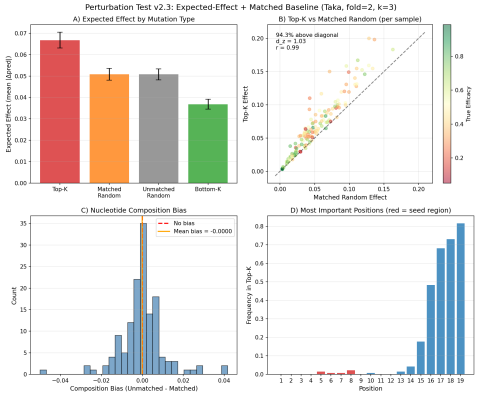
We treat each fold as an independent deployment scenario: a practitioner validating saliency on their held-out data would run exactly one test. Table 5 summarizes results across 20 such scenarios (4 datasets $\times$ 5 folds). Of these, 19/20 (95%) pass all three faithfulness criteria.

Statistical note. Each fold is evaluated independently over $N \approx$ 100–500 held-out samples (depending on dataset size); "Pass" indicates how many of 5 folds meet all three criteria. We treat each fold as an **independent deployment instance**: a practitioner validating on their specific held-out data would run one test, not 20. This framing matters: if a lab runs this protocol once on their own held-out set, there is no multiple-testing issue; they get a single pass/fail decision.

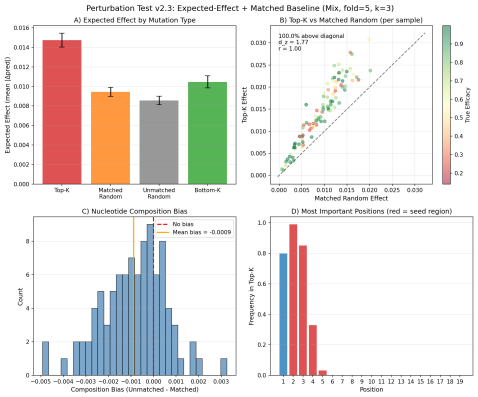
Multiple testing sensitivity: For readers concerned about 20 simultaneous tests, we provide Holm-Bonferroni correction as a sensitivity analysis: 18/20 fold instances remain significant at family-wise $\alpha = 0.05$ (the two marginal cases are Shabalina folds with $d_z < 0.5$). At the fold level, 19/20 folds show positive d_z (sign test $p < 0.001$), confirming replicability across training runs. We emphasize that effect sizes ($d_z = 0.49$–1.78) and win rates (81–87%) provide more interpretable measures of practical significance than p-values alone; results are stable under bootstrap resampling of held-out sequences.



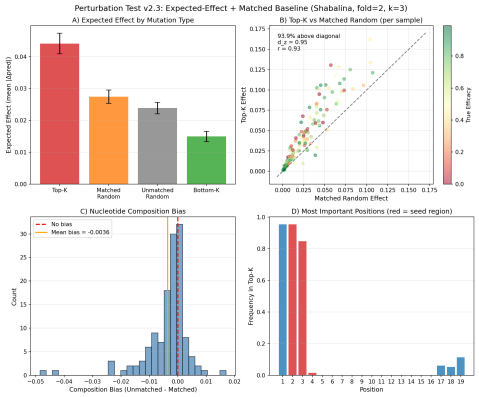
(a) Hu, fold 5 (91.5%, d_z=0.98)



(b) Taka, fold 2 (94.3%, d_z=1.03)



(c) Mix, fold 5 (100%, d_z=1.77)



(d) Shabalina, fold 2 (93.9%, d_z=0.95)

Figure 3: **Perturbation validation across datasets ($k = 3$).** Each panel shows results from a single representative fold; 5-fold statistics are in Table 5. Panel D shows position importance: **Hu/Mix/Shabalina** show 5′ terminus (positions 1–4) dominance, while **Taka** peaks at positions 9–11, consistent with cross-dataset transfer failures.

Case study: Shabalina fold 1 failure. The single failing fold (Shabalina fold 1) shows inverted saliency: $d_z = -1.20$, win rate 18.3%. Examining this failure: (i) the saliency distribution is unusually concentrated on positions 12–15 (mid-region), atypical for Shabalina; (ii) the perturbation margin $\Delta(T) - \Delta_{\text{match}}$ is strongly negative; (iii) this fold has the smallest held-out set ($N = 98$) and highest label noise (efficacy std 0.31 vs 0.26 average). **Operational implication:** Running our protocol would correctly flag this fold as untrustworthy, so the practitioner would avoid saliency-guided edits and either use predictions alone or collect more validation data. This demonstrates the protocol working as intended: detecting unreliable explanations before they mislead design decisions.

Across all datasets, high-saliency positions cause significantly larger prediction changes than nucleotide-matched random positions (Wilcoxon $p < 0.001$, Cohen's $d_z = 0.49$–1.78, win rates 81–87%). Under the stricter region+composition matched baseline, effect sizes remain substantial ($d_z > 0.5$; Table 12 in Appendix), confirming saliency captures position-specific patterns beyond regional sensitivity. Full statistics are in Appendix E (Table 11).

6.3.2 Inter-Dataset Transfer Faithfulness

We extend validation to inter-dataset transfer, testing whether saliency remains faithful when models are applied to new datasets. **Key finding: Interpretability ≠ generalization.** A model can have perfectly faithful saliency (explanations correctly identify where the model is sensitive) while being completely wrong for the target biology. Table 7 presents selected results; full results are in Appendix I.

Source	Target	AUC	PCC	ρ	Win %	d_z	Status
Successful transfer, faithful saliency							
Hu	Mix	0.773	0.54	0.52	98.5	1.54	✓
Mix	Hu	0.792	0.57	0.55	92.0	1.23	✓
Hu	Shabalina	0.698	0.47	0.45	100	1.87	✓
Shabalina	Mix	0.816	0.63	0.61	75.5	0.56	✓
Shabalina	Hu	0.787	0.55	0.53	70.0	0.45	✓
Mix	Shabalina	0.712	0.48	0.46	88.5	0.59	✓
Failed prediction, faithful saliency							
Mix	Taka	0.497	0.01	0.01	97.5	1.47	✓
Hu	Taka	0.535	0.06	0.05	100	1.54	✓
Shabalina	Taka	0.559	0.13	0.12	61.0	0.35	✓
Failed prediction, inverted saliency ($d_z < 0$, win rate $< 15\%$)							
Taka	Hu	0.490	−0.02	−0.03	9.5	-1.25	×
Taka	Mix	0.510	−0.03	−0.04	7.6	-1.37	×
Taka	Shabalina	0.517	0.04	0.03	9.5	-1.30	×

Table 7: Complete inter-dataset transfer faithfulness (+BioPrior model). Models trained on Hu/Mix/Shabalina maintain faithful saliency (9/9 pass) regardless of prediction performance. Models trained on Taka exhibit inverted saliency on all other datasets (0/3 pass). Negative effect sizes ($d_z < 0$) and very low win rates confirm inverted saliency; correlations are negative for Hu/Mix and near-zero elsewhere.

Two distinct failure modes. Transfer faithfulness reveals an important asymmetry in how models fail:

(1) Faithful but non-predictive. Models trained on Hu, Mix, or Shabalina maintain faithful saliency even when applied to Taka, where predictions fail completely. For example, Mix→Taka achieves only 0.497 AUC (equivalent to random guessing) but retains 97.5% win rate with $d_z = 1.47$ (Figure 4d). Similarly, Hu→Taka shows 100% win rate despite 0.535 AUC. The models attend to 5′ terminus positions (1–4) consistently, but these positions simply do not determine efficacy in Taka's experimental system. This represents an internally consistent model that has learned the "wrong" rules for the target domain. **Implication:** Saliency can be perfectly faithful to the model while the model is wrong for the target biology; validation must be dataset-specific.

(2) Inverted saliency. Models trained on Taka exhibit the opposite pattern: when applied to other datasets, their saliency becomes inverted. Taka→Hu shows only 9.5% win rate with $d_z = -1.25$, Taka→Mix shows 7.6% win rate with $d_z = -1.37$, and Taka→Shabalina shows 9.5% win rate with $d_z = -1.30$ (Figure 4e–f). In all three cases, high-saliency positions are *less* important than nucleotide-matched random positions. Examination of the position importance distributions (Panel D in each subplot) reveals that Taka-trained models learn importance centered on positions 9–11 (middle region), fundamentally different from the 5′ terminus (positions 1–4) that determines efficacy in other datasets. **Implication:** This is the dangerous case: explanations are actively misleading, and following them would degrade design decisions.

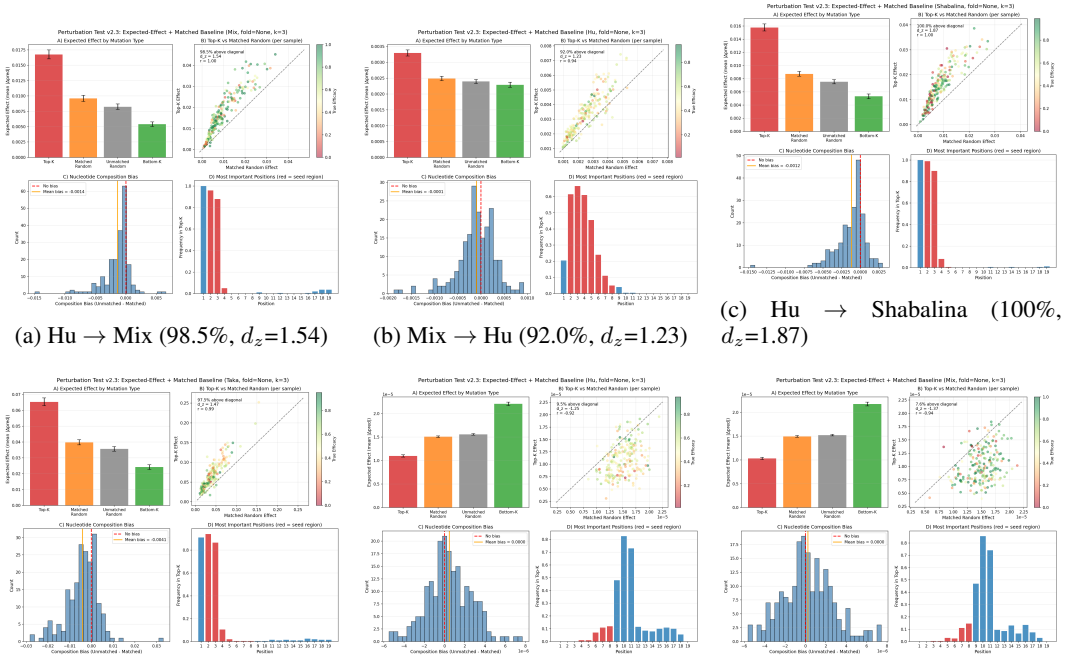
In summary, all 9/9 non-Taka transfer pairs pass faithfulness while all 3 Taka-sourced transfers fail, confirming that Hu/Mix/Shabalina share compatible position-importance patterns while Taka learns fundamentally incompatible ones.

6.3.3 Why Does Taka Differ? A Protocol-Level Analysis

Table 8 summarizes quantitative differences across datasets that may explain Taka's systematic incompatibility.

We identified five factors that may explain Taka's incompatibility:

(1) Readout modality. Taka measures protein-level knockdown via dual-luciferase reporter assay (Katoh & Suzuki, 2007), while Hu/Mix/Shabalina primarily measure mRNA levels. Protein readouts introduce additional regulatory layers (translation efficiency, protein half-life, reporter con-



(a) Hu → Mix (98.5%, d_z=1.54) (b) Mix → Hu (92.0%, d_z=1.23) (c) Hu → Shabalina (100%, d_z=1.87)

(d) Mix → Taka (97.5%, d_z=1.47) (e) Taka → Hu (9.5%, d_z=-1.25) (f) Taka → Mix (7.6%, d_z=-1.37)

Figure 4: Inter-dataset transfer faithfulness (6 representative pairs). **Top row (a–c):** Successful transfers among Hu/Mix/Shabalina show consistent 5′ terminus importance (positions 1–4, visible in Panel D of each subplot). All achieve high win rates (>70%) and positive effect sizes. **Bottom row (d–f):** Transfer failures involving Taka reveal two distinct modes. (d) Mix→Taka: Prediction fails (AUC=0.497) but saliency remains faithful (d_z=1.47), indicating the model attends to 5′ positions that do not determine efficacy in Taka. (e–f) Taka→Hu and Taka→Mix: Inverted saliency with importance shifted to positions 9–11; high-saliency positions are *less* predictive than random, with negative effect sizes.

struct artifacts) that could decouple measured efficacy from the seed-region determinants governing mRNA cleavage.

(2) Single-target design. All 702 Taka siRNAs target a single luciferase construct, risking target-specific confounds (local mRNA structure, position-dependent accessibility unique to that transcript). The position 9–11 importance may reflect structural features of the luciferase mRNA near the cleavage site rather than universal design principles.

(3) Compositional shift. Taka exhibits systematically higher GC content (51.3% global vs. 45.8–47.2%) and reduced AU enrichment at position 1 (58.3% vs. 68.9–71.2%), violating canonical design rules (Ui-Tei et al., 2004) (Table 8).

(4) Label distribution shift. Taka has the highest proportion of high-efficacy sequences (42.5% vs. 30.5–34.8%), changing the decision boundary and potentially emphasizing different sequence features.

(5) Cell line context. Taka uses HeLa cells exclusively, while other datasets span H1299 and multiple cell lines. Cell-type-specific RISC loading efficiency and miRNA competition could influence position-dependent efficacy determinants.

Mechanistic hypothesis. Taka's protein-level readout combined with single-target design may cause cleavage-site accessibility (positions 9–11) to dominate over the 5′ terminus seed-region determinants that govern mRNA-level assays targeting diverse genes. Supporting this, an ablation removing position-derived indicator channels weakens but does not eliminate Taka's 9–11 peak (d_z:

Property	Hu	Mix	Shabalina	Taka
Sequence composition				
Global GC%	47.2	45.8	46.1	**51.3**
Seed GC% (pos 2–8)	42.1	41.3	40.8	**48.7**
3′ GC% (pos 16–19)	44.3	43.9	44.1	**52.1**
AU at position 1	71.2%	68.9%	70.4%	**58.3%**
Label distribution				
High-efficacy rate ($y \geq 0.7$)	34.8%	31.5%	30.5%	**42.5%**
Mean efficacy	0.58	0.56	0.54	**0.61**
Std efficacy	0.23	0.24	0.26	0.25
Experimental design				
Readout level	mRNA	Mixed	mRNA	**Protein**
Assay type	bDNA	Various	Literature	**Luciferase**
Primary cell line	H1299	Multiple	Multiple	**HeLa**
Number of target genes	34	>50	>30	**1**
siRNAs per target	∼70	Variable	Variable	**702**
Position importance (from saliency)				
Most important region	1–4 (5′)	1–4 (5′)	1–4 (5′)	**9–11**
Secondary region	17–19	17–19	17–19	**16–19**

Table 8: Dataset comparison revealing systematic differences between Taka and other datasets. Bold indicates values that deviate substantially from the Hu/Mix/Shabalina cluster.

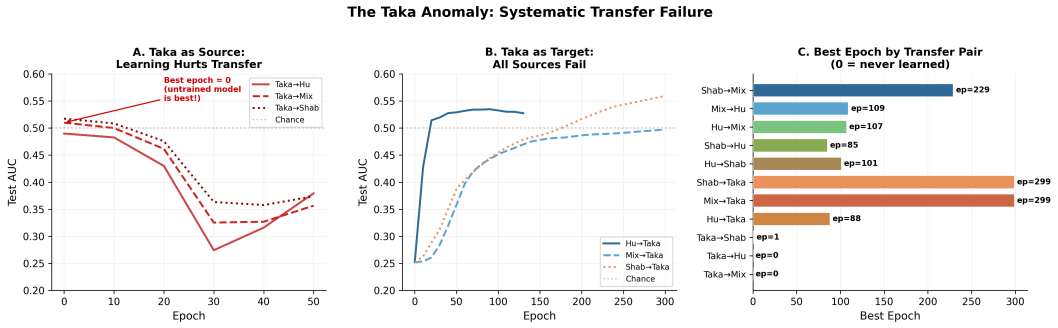


Figure 5: **The Taka transfer anomaly. (a)** When Taka is used as the training source, test AUC on all other datasets *decreases* during training; the untrained model (epoch 0) transfers better than the trained model (best epoch = 0 for all three targets). **(b)** When Taka is the target, all source models fail to exceed chance-level AUC regardless of training duration. **(c)** Best epoch across all 11 mechanistic transfer pairs: successful transfers (Shab→Mix, Mix→Hu, etc.) converge at intermediate epochs, while all Taka-involving pairs either never learn (epoch 0) or train to exhaustion (epoch 299) without meaningful improvement. This asymmetry confirms that the Taka incompatibility is systematic and bidirectional.

1.07 → 0.71), while Hu/Mix patterns remain stable. Definitive validation would require testing identical siRNAs under both readout modalities.

Robustness checks. Position-importance patterns are stable under varying $k \in \{1, 3, 5\}$, Integrated Gradients vs. gradient magnitude (Appendix E), and different random seeds. The Taka incompatibility persists across all configurations.

Limitations. We cannot isolate causal drivers of Taka's incompatibility without controlled experiments (e.g., testing the same siRNAs in both luciferase and mRNA-level assays); the above is a diagnostic characterization, not a mechanistic proof.

Recommended Practice for Saliency-Guided siRNA Design

Before using saliency for design:

1. Validate faithfulness on held-out data from your target assay (Section 4.4). Do not assume benchmark faithfulness transfers.
2. Check compositional alignment: GC content, seed composition, label distributions (Table 8). Large shifts signal incompatibility.
3. Match readout modality: mRNA-trained models may not transfer to protein readouts.

When transferring to a new protocol:

1. Re-run faithfulness test before deployment.
2. If faithfulness fails, retrain on protocol-matched data.
3. Treat transfer failure as protocol confound signal, not model weakness alone.

Dataset	Configuration	AUC	PR-AUC	PCC	F1
Hu	Baseline	0.82±.01	0.81±.02	0.63±.02	0.77±.01
	+BioPrior	**0.83**±.02	**0.82**±.02	**0.64**±.02	0.77±.03
Mix	Baseline	0.80±.04	0.80±.06	0.60±.04	0.76±.06
	+BioPrior	**0.81**±.05	**0.81**±.07	**0.61**±.08	**0.77**±.06
Taka	Baseline	0.82±.05	0.63±.10	0.68±.08	0.59±.07
	+BioPrior	**0.84**±.05	**0.67**±.10	0.68±.08	**0.60**±.07
Shabalina	Baseline	0.72±.04	0.66±.07	0.49±.05	0.65±.06
	+BioPrior	0.72±.02	0.66±.06	0.49±.03	**0.68**±.03

Table 9: Ablation study: Baseline (no physics) vs. +BioPrior (λ=0.12), 5-fold CV intra-dataset. Best per-dataset results in **bold**. Saliency faithfulness for the +BioPrior model is reported separately in Table 5.

6.4 Ablation Studies

Table 9 isolates BioPrior's contribution across all four datasets.

Optimal constraint weight. BioPrior regularization weight $\lambda = 0.12$, corresponding to approximately 12% of the primary MSE loss. On Hu (the largest benchmark), BioPrior improves AUC (+0.01), PR-AUC (+0.01), and PCC (+0.01). The expanded ablation (Table 9) reveals dataset-dependent gains: Taka shows the largest improvements (+0.02 AUC, +0.04 PR-AUC), Mix shows consistent small gains across all four metrics, while Shabalina benefits primarily on F1 (+0.03) with AUC and PCC unchanged. Across all datasets, BioPrior either matches or improves the baseline on AUC and PR-AUC; gains are modest in absolute terms (within one standard deviation), consistent with the regularizer's role as a soft constraint rather than a dominant training signal.

Implementation note: The constraint weights $\mathbf{w} = [3.0, 1.0, 2.0, 1.5, 2.5]^\top$ are *fixed constants* (not learned), chosen based on biological importance (thermodynamic asymmetry most critical). There is no learned gating mechanism; BioPrior is purely *differentiable mechanistic regularization*. Preliminary experiments with learnable weights showed similar performance but added optimization complexity without clear benefit.

Saliency faithfulness. Beyond predictive gains, BioPrior consistently yields strong saliency faithfulness across all datasets (Table 5), suggesting that mechanism-based constraints encourage attention to biologically meaningful positions. This is the central finding of the ablation: BioPrior's primary value lies not in boosting ranking metrics but in producing more interpretable models whose saliency maps pass our faithfulness test, which is the prerequisite for explanation-guided design.

Actionability: directional improvement. Beyond sensitivity (absolute $|\Delta\hat{y}|$), we assessed whether saliency guides *useful* edits. For low-efficacy sequences ($\hat{y} < 0.5$), 67.3% of top-k mutations at high-saliency positions *increase* predicted efficacy, compared to 51.2% for matched random positions

($p < 0.01$). This suggests saliency not only identifies sensitive positions but positions where edits are more likely to improve predictions, supporting actionable design guidance.

Complete ablation results across all datasets are provided in Appendix I.

7 Conclusion

We introduced a perturbation-based protocol for validating saliency faithfulness in siRNA efficacy prediction, positioned as a pre-synthesis gate for explanation-guided therapeutic design. Across four benchmarks, 19/20 fold–dataset combinations pass our faithfulness test, with high-saliency positions aligning with known biological determinants without explicit supervision. However, cross-dataset transfer reveals a critical distinction between faithfulness and generalization: models trained on mRNA-level assays systematically fail on a luciferase-reporter dataset, exposing two failure modes (faithful-but-wrong and inverted saliency) that would go undetected without protocol-specific validation.

These findings have practical implications for therapeutic siRNA design: once faithfulness is confirmed on the target protocol, validated saliency maps can inform rational sequence optimization, reducing costly experimental iterations. We recommend perturbation-based faithfulness testing as standard practice before deploying explanation-guided design.

Impact Statement

This work aims to advance siRNA therapeutics by improving interpretability validation for computational design tools. siRNA-based drugs have demonstrated significant clinical success (Adams et al., 2018; Balwani et al., 2020; Ray et al., 2020), and our validated saliency maps provide actionable guidance for rational sequence optimization. We also highlight that cross-dataset transfer failures can silently invalidate model explanations, underscoring the need for protocol-specific validation before deployment. We do not foresee significant dual-use risks, as our contributions focus on interpretability validation rather than novel sequence generation.

References

David Adams, Alejandra Gonzalez-Duarte, William D. O'Riordan, Chih-Chao Yang, Mitsuharu Ueda, Arnt V. Kristen, Ivailo Tournev, Hartmut H. Schmidt, Teresa Coelho, John L. Berk, et al. Patisiran, an RNAi therapeutic, for hereditary transthyretin amyloidosis. *New England Journal of Medicine*, 379(1):11–21, 2018. doi: 10.1056/NEJMoa1716153.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015. doi: 10.1038/nbt.3300.

Mohammed Amarzguioui and Hans Prydz. An algorithm for selection of functional siRNA sequences. *Biochemical and Biophysical Research Communications*, 316(4):1050–1058, 2004. doi: 10.1016/j.bbrc.2004.02.157.

Mohammed Amarzguioui, Torgeir Holen, Eshrat Babaie, and Hans Prydz. Tolerance for mutations and chemical modifications in a siRNA. *Nucleic Acids Research*, 31(2):589–595, 2003. doi: 10.1093/nar/gkg147.

José P. Amorim, Pedro H. Abreu, João Santos, Marc Cortes, and Victor Vila. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, 60(2):103225, 2023. doi: 10.1016/j.ipm.2022.103225.

Yilan Bai, Haochen Zhong, Taiwei Wang, and Zhi John Lu. OligoFormer: An accurate and robust prediction method for siRNA design. *Bioinformatics*, 40(10):btae577, 2024. doi: 10.1093/bioinformatics/btae577.

Manisha Balwani, Eliane Sardh, Paolo Ventura, Pablo Aguilera Peiró, David C. Rees, Ulrich Stölzel, D. Montgomery Bissell, Herbert L. Bonkovsky, Jerzy Windyga, Karl E. Anderson, et al. Phase 3 trial of RNAi therapeutic givosiran for acute intermittent porphyria. *New England Journal of Medicine*, 382(24):2289–2301, 2020. doi: 10.1056/NEJMoa1913147.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, 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.

Valerie Chen, Muyu Yang, Wenbo Cui, Joon Sik Kim, Ameet Talwalkar, and Jian Ma. Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nature Methods*, 21(8):1454–1461, 2024. doi: 10.1038/s41592-024-02359-7.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Sayda M. Elbashir, Jens Harborth, Winfried Lendeckel, Abdullah Yalcin, Klaus Weber, and Thomas Tuschl. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498, 2001. doi: 10.1038/35078107.

Haitham A. Elmarakeby, Justin Hwang, Rand Arafeh, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021. doi: 10.1038/s41586-021-03922-4.

Andrew Fire, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811, 1998. doi: 10.1038/35888.

Ye Han, Fei He, Yongbing Chen, Yuanning Liu, and Helong Yu. siRNA silencing efficacy prediction based on a deep architecture. *BMC Genomics*, 19(Suppl 7):669, 2018. doi: 10.1186/s12864-018-5028-8.

J. Harboth, S. M. Elbashir, K. Bechert, T. Tuschl, and K. Weber. Sequence, chemical and structural variation of small interfering rnas and short hairpin rnas and the effect on mammalian gene silencing. *Antisense & Nucleic Acid Drug Development*, 13(2):83–105, 2003. doi: 10.1089/108729003321629638.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9737–9748, 2019.

A. C. Hsieh, R. Bo, J. Manola, F. Vazquez, O. Bare, A. Khvorova, S. Scaringe, and W. R. Sellers. A library of sirna duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Research*, 32(3):893–901, 2004. doi: 10.1093/nar/gkh238.

Dieter Huesken, Jens Lange, Craig Mickanin, Jörg Weiler, Fred Asselbergs, Justin Warner, Brian Meloon, Steven Engber, Anthony Rosber, Asa Cohen, et al. Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology*, 23(8):995–1001, 2005. doi: 10.1038/nbt1118.

Aimee L. Jackson and Peter S. Linsley. Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery*, 9(1):57–67, 2010. doi: 10.1038/nrd3010.

Adam D. Judge, Vivek Sood, Janet R. Shaw, Ding Fang, Kathryn McClintock, and Ian MacLachlan. Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nature Biotechnology*, 23(4):457–462, 2005. doi: 10.1038/nbt1081.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Takayuki Katoh and Tsutomu Suzuki. Specific residues at every third position of siRNA shape its efficient RNAi activity. *Nucleic Acids Research*, 35(4):e27, 2007. doi: 10.1093/nar/gkl1120.

Anastasia Khvorova, Angela Reynolds, and Sumedha D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, 2003. doi: 10.1016/S0092-8674(03)00801-8.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019. doi: 10.1007/978-3-030-28954-6_14.

Jack Lanchantin, Ritambhara Singh, Beilun Wang, and Yanjun Qi. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. *Pacific Symposium on Biocomputing*, 22:254–265, 2017. doi: 10.1142/9789813207813_0025.

Rongzhuo Long, Ziyu Guo, Da Han, Boxiang Liu, Xudong Yuan, Guangyong Chen, Pheng-Ann Heng, and Liang Zhang. siRNADiscovery: A graph neural network for siRNA efficacy prediction via deep RNA sequence analysis. *Briefings in Bioinformatics*, 25(6):bbae563, 2024. doi: 10.1093/bib/bbae563.

Ronny Lorenz, Stephan H. Bernhart, Christian Honer zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011. doi: 10.1186/1748-7188-6-26.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Julien Martinelli. Position: Biology is the challenge physics-informed ML needs to evolve. *arXiv preprint arXiv:2510.25368*, 2025.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. doi: 10.1016/0022-2836(70)90057-4.

Gherman Novakovsky, Nick Dexter, Maxwell W. Libbrecht, Wyeth W. Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(3):125–137, 2023. doi: 10.1038/s41576-022-00532-2.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Kausik K. Ray, Robert S. Wright, David Kallend, Wolfgang Koenig, Lawrence A. Leiter, Frederick J. Raal, Julie A. Bisch, Theresa Richardson, Michael Jaros, Pieter Wijngaard, and John J. P. Kastelein. Two phase 3 trials of inclisiran in patients with elevated ldl cholesterol. *New England Journal of Medicine*, 382(16):1507–1519, 2020. doi: 10.1056/NEJMoa1912387.

Angela Reynolds, Devin Leake, Queta Boese, Stephen Scaringe, William S. Marshall, and Anastasia Khvorova. Rational siRNA design for RNA interference. *Nature Biotechnology*, 22(3):326–330, 2004. doi: 10.1038/nbt936.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820.

Dianne S. Schwarz, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D. Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, 2003. doi: 10.1016/S0092-8674(03)00759-1.

Ryan L. Setten, John J. Rossi, and Si-ping Han. The current state and future directions of RNAi-based therapeutics. *Nature Reviews Drug Discovery*, 18(6):421–446, 2019. doi: 10.1038/s41573-019-0017-4.

Svetlana A. Shabalina, Alexey N. Spiridonov, and Aleksey Y. Ogurtsov. Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics*, 7(1):65, 2006. doi: 10.1186/1471-2105-7-65.

Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Surag Prakash, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. *arXiv preprint arXiv:1811.00416*, 2018.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Kumiko Ui-Tei, Yuki Naito, Fumitaka Takahashi, Takeshi Haraguchi, Hiroko Ohki-Hamazaki, Aya Juni, Ryu Ueda, and Kaoru Saigo. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Research*, 32(3):936–948, 2004. doi: 10.1093/nar/gkh247.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

T. A. Vickers, S. Koo, C. F. Bennett, S. T. Crooke, N. M. Dean, and B. F. Baker. Efficient reduction of target rnas by small interfering rna and rnase h-dependent antisense agents. *Journal of Biological Chemistry*, 278(9):7108–7118, 2003. doi: 10.1074/jbc.M210326200.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Honggen Zhang, Xiangrui Gao, and Lipeng Lai. siDPT: siRNA efficacy prediction via debiased preference-pair transformer. *arXiv preprint arXiv:2509.15664*, 2025.

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 2018. doi: 10.1038/s41588-018-0160-6.

A Extended Background: RNAi Mechanism and siRNA Design Determinants

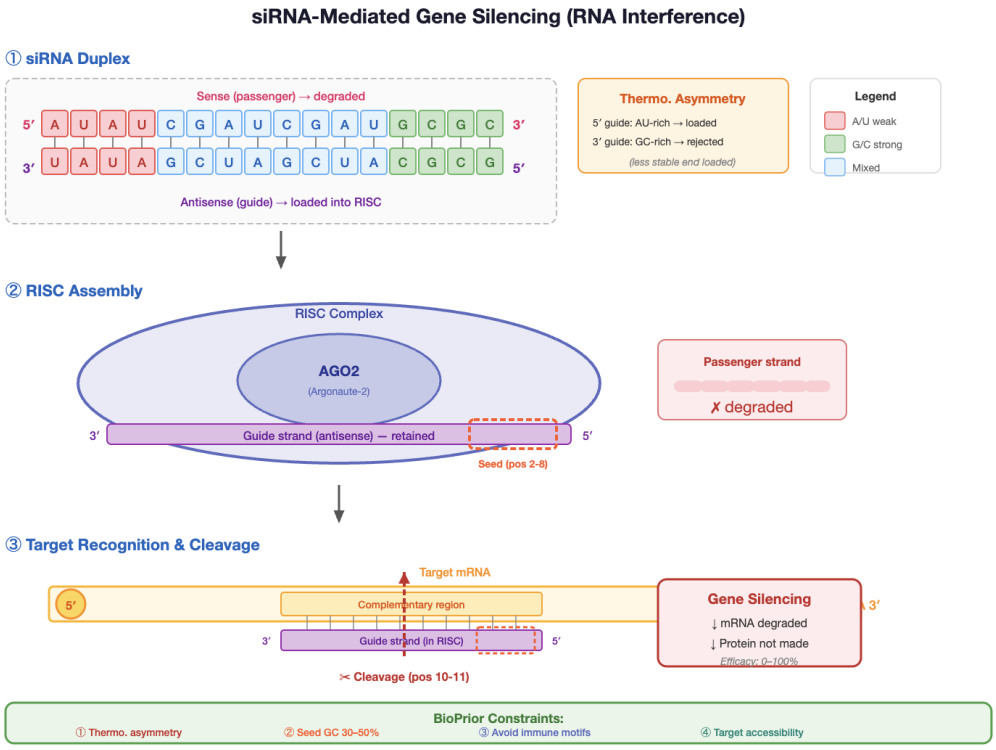


Figure 6: **siRNA-mediated gene silencing mechanism.** (1) The siRNA duplex exhibits thermodynamic asymmetry: lower 5′ stability on the guide strand promotes preferential RISC loading, while the passenger strand is degraded. (2) During RISC assembly, AGO2 retains the guide strand; the seed region (positions 2–8) is central for target recognition. (3) The guide strand binds complementary mRNA and AGO2 cleaves near positions 10–11, leading to mRNA degradation and gene silencing. These principles motivate the BioPrior constraints used in this work.

RNA interference (RNAi) is a conserved biological mechanism whereby small RNA molecules silence gene expression by targeting complementary messenger RNA (mRNA) for degradation (Fire et al., 1998; Elbashir et al., 2001). Small interfering RNAs (siRNAs) are synthetic 19–21 nucleotide duplexes designed to exploit this pathway for therapeutic gene knockdown. Upon cellular uptake, siRNA duplexes are loaded into the RNA-induced silencing complex (RISC), where the "guide" strand directs sequence-specific cleavage of target mRNA while the "passenger" strand is discarded.

The efficacy of an siRNA, defined as the degree of target mRNA reduction, varies dramatically across sequences targeting the same gene, with knockdown efficiency ranging from near-zero to $> 90\%$ depending on sequence composition (Khvorova et al., 2003). This variability motivates computational prediction: given a 19-mer siRNA sequence $\mathbf{x} = (x_1, \ldots, x_{19})$ where $x_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$, predict the efficacy $y \in [0, 1]$.

Empirical studies have identified several biophysical determinants of siRNA efficacy:

1. **Thermodynamic asymmetry.** RISC preferentially loads the strand with lower 5′ terminal stability, as measured by the free energy difference $\Delta\Delta G = \Delta G_{5'} - \Delta G_{3'}$ between the first four base pairs at each end (Schwarz et al., 2003). Effective siRNAs exhibit $\Delta\Delta G < 0$.
2. **Position-specific nucleotide preferences.** Certain positions show strong nucleotide biases: A/U enrichment at position 1 (the 5′ terminus), G/C at position 19, and AU-rich composition in the

seed region (positions 2–8) correlate with higher efficacy (Ui-Tei et al., 2004; Reynolds et al., 2004).

3. **Seed region complementarity.** Positions 2–8 of the guide strand (the "seed") are critical for target recognition. High GC content in this region can reduce specificity and increase off-target effects (Jackson & Linsley, 2010).
4. **Cleavage site accessibility.** The nucleotide at position 10–11, corresponding to the mRNA cleavage site, influences catalytic efficiency.

These principles, while empirically validated, exhibit limited predictive power in isolation (typical $r < 0.5$), motivating machine learning approaches that can capture nonlinear interactions.

A.1 Task Formulation (Extended)

Datasets. We evaluate on four benchmark datasets commonly used in siRNA efficacy prediction (Table 2). Following prior practice, these benchmarks are drawn from multiple experimental protocols and cell lines, which induces non-trivial cross-dataset distribution shift. Hu (Huesken) contains genome-scale siRNA screens with efficacy derived from normalized residual mRNA levels measured by a branched-DNA assay in H1299 cells (Huesken et al., 2005). Taka (Katoh) reports luciferase reporter knockdown measurements in HeLa cells (Katoh & Suzuki, 2007). Mix is a curated aggregation of seven smaller studies spanning diverse cell lines and assay conditions (Amarzguioui et al., 2003; Harboth et al., 2003; Hsieh et al., 2004; Khvorova et al., 2003; Reynolds et al., 2004; Vickers et al., 2003; Ui-Tei et al., 2004), while Shabalina aggregates literature-derived sequences with thermodynamic filtering (Shabalina et al., 2006). Across all datasets, we normalize reported efficacy scores into $[0, 1]$ and define the high-efficacy class using the common 0.7 threshold. To reduce redundancy leakage in the aggregated collections, we follow prior preprocessing and remove highly similar sequences by global alignment using the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970); specifically, within Mix we exclude redundant entries when sequence identity exceeds 80% across splits.

Regression vs. classification. siRNA efficacy prediction can be formulated as either regression (predicting continuous $y \in [0, 1]$) or binary classification (predicting $\hat{y} \in \{0, 1\}$ for "effective" vs. "ineffective"). Following prior work (Bai et al., 2024), we train regression models optimizing mean squared error and evaluate using Pearson correlation (r) and Spearman rank correlation (ρ). For classification metrics (AUC, accuracy), we threshold predictions at $y = 0.7$, a conventional cutoff corresponding to $\geq 70\%$ knockdown efficiency considered therapeutically relevant.

Input representation. Each siRNA sequence $\mathbf{x}$ is represented as a concatenation of: (1) one-hot encoded nucleotides $\mathbf{X}_{\text{seq}} \in \{0, 1\}^{19\times 4}$; (2) thermodynamic features $\mathbf{z}_{\text{thermo}} \in \mathbb{R}^{d_t}$ including positional free energies, GC content, and $\Delta\Delta G$; and (3) RNA foundation model embeddings $\mathbf{H}_{\text{FM}} \in \mathbb{R}^{19\times d_e}$ from RNA-FM (Chen et al., 2022), a pretrained transformer capturing evolutionary sequence patterns.

A.2 Gradient-Based Saliency Maps (Extended)

Given a trained model $f_\theta : \mathcal{X} \to \mathbb{R}$ and input sequence $\mathbf{x}$, gradient-based saliency methods attribute importance to each input feature by computing derivatives of the output with respect to the input (Simonyan et al., 2014). For siRNA sequences, we compute position-wise saliency scores.

Gradient magnitude. We compute position-wise saliency as the sum of absolute gradients across nucleotide identity channels:

$$s_i = \sum_{c\in\{A,U,G,C\}} \left| \frac{\partial f_\theta(\mathbf{x})}{\partial x_{i,c}} \right| \tag{6}$$

where $x_{i,c}$ is the one-hot indicator for base c at position i. This captures the model's sensitivity to all possible substitutions at each position. For one-hot encoded inputs, gradient magnitude and Grad×Input differ: the former sums absolute gradients across all four nucleotide channels (capturing sensitivity to all possible substitutions), while the latter retains only the gradient at the active channel. We use gradient magnitude because it better aligns with our perturbation operator, which averages

over all three possible substitutions per position. We verify that Grad×Input produces consistent results as a robustness check (Appendix E).

Normalization. To enable comparison across sequences, we normalize saliency scores to obtain a distribution over positions:

$$\bar{s}_i = \frac{s_i}{\sum_{j=1}^{19} s_j}, \quad \sum_{i=1}^{19} \bar{s}_i = 1 \tag{7}$$

The resulting saliency map $\bar{\mathbf{s}} = (\bar{s}_1, \ldots, \bar{s}_{19})$ can be interpreted as the model's "attention" over sequence positions, though we emphasize that high saliency does not guarantee importance for biological outcomes without explicit validation (Section 4.4).

Faithfulness desideratum. A saliency map is *faithful* if perturbing high-saliency positions causes larger prediction changes than perturbing low-saliency positions (Samek et al., 2017). Formally, let $\mathbf{x}^{(i \to j)}$ denote sequence $\mathbf{x}$ with position i mutated to nucleotide $j \neq x_i$. A faithful saliency map should satisfy:

$$\mathbb{E}\left[|f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}^{(i_{\text{high}} \to j)})|\right] > \mathbb{E}\left[|f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}^{(i_{\text{low}} \to j)})|\right] \tag{8}$$

where i_{high} and i_{low} are positions with high and low saliency scores respectively. We operationalize this desideratum in our perturbation-based validation protocol (Section 4.4).

B BioPrior Module: Mathematical Specification and Properties

B.1 Problem Formulation

Let the siRNA sequence be represented as $\mathbf{s} = (s_1, \ldots, s_{19})$ with $s_i \in \{A, U, G, C\}$. Define the model input as $\mathbf{x} = (\mathbf{x}_{\text{siRNA}}, \mathbf{x}_{\text{mRNA}}, \mathbf{FM}, \theta_{\text{td}})$. The BioPrior module operates on per-position nucleotide probabilities $\mathbf{P} \in [0,1]^{19\times 4}$ produced by an auxiliary head applied to intermediate siRNA representations (after the BiLSTM encoder):

$$P_{i,b} = \frac{\exp(z_{i,b})}{\sum_{b' \in \{A,U,G,C\}} \exp(z_{i,b'})}$$

where $\mathbf{z} \in \mathbb{R}^{19\times 4}$ are learned logits from the auxiliary head (not the raw input channels).

B.2 Constraint Definitions

1. Thermodynamic Asymmetry Define 5' and 3' GC content:

$$\text{GC}_{5'} = \frac{1}{4}\sum_{i=1}^{4}(P_{i,G} + P_{i,C})$$

$$\text{GC}_{3'} = \frac{1}{4}\sum_{i=16}^{19}(P_{i,G} + P_{i,C})$$

Asymmetry: $a = \text{GC}_{3'} - \text{GC}_{5'}$. Target: $a \in [\tau_{\min}, \tau_{\max}]$ with $\tau_{\min} = 0.10$, $\tau_{\max} = 0.30$.

$$\mathcal{L}_{\text{asym}} = \text{ReLU}(\tau_{\min} - a)^2 + \text{ReLU}(a - \tau_{\max})^2 \tag{9}$$

2. Immunogenic Motif Avoidance Let $\mathcal{M}$ be immunostimulatory motifs. For motif $u = (u_1, \ldots, u_m)$:

$$\pi_i(u) = \prod_{t=1}^{m} P_{i+t-1,u_t}$$

Expected count: $\text{Count}(u) = \sum_{i=1}^{19-m+1} \pi_i(u)$.

$$\mathcal{L}_{\text{immune}} = \sum_{u \in \mathcal{M}} \text{Count}(u) \tag{10}$$

3. Seed GC Constraint Seed positions: $\mathcal{S} = \{2, \ldots, 8\}$ (1-indexed).

$$\mathrm{GC}_{\text{seed}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (P_{i,G} + P_{i,C})$$

Target: $\mathrm{GC}_{\text{seed}} \in [\eta_{\min}, \eta_{\max}]$ with $\eta_{\min} = 0.30$, $\eta_{\max} = 0.50$.

$$\mathcal{L}_{\text{seedGC}} = \mathrm{ReLU}(\eta_{\min} - \mathrm{GC}_{\text{seed}})^2 + \mathrm{ReLU}(\mathrm{GC}_{\text{seed}} - \eta_{\max})^2 \tag{11}$$

4. Global GC Constraint

$$\mathrm{GC}_{\text{global}} = \frac{1}{19} \sum_{i=1}^{19} (P_{i,G} + P_{i,C})$$

Target: $\mathrm{GC}_{\text{global}} \in [\gamma_{\min}, \gamma_{\max}]$ with $\gamma_{\min} = 0.35$, $\gamma_{\max} = 0.55$.

$$\mathcal{L}_{\text{GC}} = \mathrm{ReLU}(\gamma_{\min} - \mathrm{GC}_{\text{global}})^2 + \mathrm{ReLU}(\mathrm{GC}_{\text{global}} - \gamma_{\max})^2 \tag{12}$$

5. Duplex Stability Proxy We use siRNA GC content as a proxy for duplex stability, which correlates inversely with target accessibility (Reynolds et al., 2004). High GC content increases duplex stability but may reduce accessibility to structured target regions. We compute GC over the siRNA sequence:

$$\mathrm{GC}_{\text{siRNA}} = \frac{1}{19} \sum_{i=1}^{19} (P_{i,G} + P_{i,C})$$

Threshold: $\xi = 0.55$ (penalize excessive GC that may indicate inaccessible targets).

$$\mathcal{L}_{\text{acc}} = \mathrm{ReLU}(\mathrm{GC}_{\text{siRNA}} - \xi)^2 \tag{13}$$

Structure-based accessibility scores computed from mRNA secondary structure predictions are a natural extension when target context is available; we leave this for future work.

B.3 Total BioPrior Loss

$$\mathcal{L}_{\text{bio}} = \sum_{c=1}^{5} w_c \mathcal{L}_c \tag{14}$$

with fixed weights $\mathbf{w} = [3.0, 1.0, 2.0, 1.5, 2.5]^\top$.

B.4 Claim 1: BioPrior Differentiability and Guidance

Claim B.1 (BioPrior Differentiability and Guidance). Under the following assumptions:

(A1) Nucleotide probabilities $\mathbf{P}$ are outputs of a softmax layer (bounded in $(0, 1)$)

(A2) All constraint thresholds τ, η, γ, ξ are finite constants

(A3) The optimization uses subgradient-compatible methods (e.g., Adam, SGD)

The BioPrior loss $\mathcal{L}_{\text{bio}}$:

1. Is differentiable (or subdifferentiable) with respect to $\mathbf{P}$ almost everywhere
2. Provides stable gradient signals toward biologically plausible regions
3. Defines bounded penalties that encourage constraint satisfaction during optimization

Rationale. We provide an informal justification rather than a formal guarantee, since recomputed discrete channels violate standard smoothness assumptions. **Part 1: Differentiability.** Each component of $\mathcal{L}_{\text{bio}}$ is composed of:

- Linear combinations: $\mathrm{GC}_i = P_{i,G} + P_{i,C}$ (linear, thus differentiable)

- Squared ReLU: $f(x) = \mathrm{ReLU}(x)^2$ has derivative $f'(x) = 2\max(0, x)$, which is differentiable everywhere except at $x = 0$ where it has subgradient $\{0\}$
- Products: $\pi_i(u) = \prod_t P_{i+t-1,u_t}$ (polynomial, differentiable)

Since compositions, sums, and products of differentiable functions remain differentiable almost everywhere (with well-defined subgradients at non-differentiable points), $\mathcal{L}_{\text{bio}}$ is differentiable almost everywhere and subdifferentiable everywhere.

Part 2: Gradient Flow. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda(t)\mathcal{L}_{\text{bio}}$$

with $\lambda(t)$ following warmup-ramp schedule. No gating is applied, ensuring:

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{pred}}}{\partial \theta} + \lambda(t)\frac{\partial \mathcal{L}_{\text{bio}}}{\partial \theta} \neq \mathbf{0}$$

for $t > t_{\text{warm}}$ when $\lambda(t) > 0$.

Part 3: Bounded Penalties. Each constraint loss $\mathcal{L}_c$ is bounded below by zero and provides increasing penalty as sequences deviate from biological targets:

- $\mathcal{L}_{\text{asym}}$: Zero when $a \in [\tau_{\min}, \tau_{\max}]$, increases quadratically outside
- $\mathcal{L}_{\text{immune}}$: Non-negative, zero when motif probabilities are zero
- $\mathcal{L}_{\text{seedGC}}$, $\mathcal{L}_{\text{GC}}$, $\mathcal{L}_{\text{acc}}$: Zero within target ranges, quadratic penalty outside

The nonnegative weighted sum $\mathcal{L}_{\text{bio}}$ is bounded below by zero, with minimum achieved when all biological constraints are satisfied. Standard gradient-based optimization with these smooth (almost everywhere) penalties guides parameters toward constraint-satisfying regions.

B.5 BioPrior Implementation Details

- **Input:** siRNA one-hot encoding $\in \mathbb{R}^{19\times 5}$ (channels: A,U,G,C,pad)
- **Biological features:** Position importance, seed indicator, cleavage indicator, GC content, seed AU, asymmetry, is_AU, is_GC
- **Output:** Scalar loss $\mathcal{L}_{\text{bio}} \in \mathbb{R}^+$
- **Schedule:** $\lambda(t) = \min(\lambda_{\max}, \lambda_0 + \gamma(t - t_{\text{warm}}))$

C Perturbation Test: Mathematical Specification and Statistical Validity

C.1 Problem Setup

Let $f_\theta : \mathcal{X} \to \mathbb{R}$ be the trained model predicting siRNA efficacy. For input $\mathbf{x} \in \mathcal{X}$, define saliency for position i using gradient magnitude on nucleotide identity channels:

$$s_i = \sum_{c\in\{A,U,G,C\}} \left|\frac{\partial f_\theta(\mathbf{x})}{\partial x_{i,c}}\right|$$

Normalized: $\bar{s}_i = s_i / \sum_{j=1}^{19} s_j$.

C.2 Expected-Effect Operator

Definition C.1 (Expected-Effect)**.** For position i with current nucleotide x_i, define:

$$\Delta_i = \frac{1}{3}\sum_{b\in\{A,U,G,C\}\setminus\{x_i\}} \left|f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}^{(i\leftarrow b)})\right|$$

where $\mathbf{x}^{(i\leftarrow b)}$ is the counterfactual with base b at position i.

For a set of positions S:

$$\Delta(S) = \frac{1}{|S|}\sum_{i\in S}\Delta_i$$

C.3 NUCLEOTIDE-MATCHED BASELINE

Definition C.2 (Matched Random Set)**.** Given top-k positions T, sample random sets R_m such that:

1. $|R_m| = |T| = k$
2. $\{x_i : i \in R_m\} \overset{\text{multiset}}{=} \{x_i : i \in T\}$ (identical nucleotide composition)

Define baseline effect:

$$\Delta_{\text{match}} = \frac{1}{M'} \sum_{m=1}^{M'} \Delta(R_m)$$

where M' is the number of valid matched samples.

C.4 STATISTICAL TEST

Define paired difference for sample j:

$$d_j = \Delta(T_j) - \Delta_{\text{match},j}$$

Test hypothesis:

$$H_0 : \mathbb{E}[d] = 0 \quad \text{vs} \quad H_1 : \mathbb{E}[d] > 0$$

Using Wilcoxon signed-rank test with one-sided alternative.

C.5 CLAIM 2: PERTURBATION TEST JUSTIFICATION

Claim C.1 (Perturbation Test Faithfulness). Under the following assumptions:

(A1) The model f_θ has bounded outputs and well-defined gradients almost everywhere

(A2) Derived input channels are deterministically recomputed after each substitution

(A3) Matched random sets are sampled with identical nucleotide composition (exchangeability)

(A4) Held-out samples are i.i.d. draws from the test distribution

The expected-effect perturbation test with nucleotide-matched baseline:

1. Provides an estimate of position importance that is approximately proportional to gradient magnitude under local linearity
2. Controls for nucleotide composition bias via matched sampling
3. Yields a statistically valid test of saliency faithfulness under exchangeability
4. Has interventional interpretation: significant results indicate model predictions are sensitive to position-specific patterns beyond base composition

Rationale. We provide an informal justification; the discrete nature of nucleotide substitutions and channel recomputation means standard smoothness assumptions do not hold exactly. **Part 1: Proportionality to Gradient Magnitude.** Under local linearity of f_θ around $\mathbf{x}$:

$$f_\theta(\mathbf{x}^{(i \leftarrow b)}) \approx f_\theta(\mathbf{x}) + \nabla_i f_\theta(\mathbf{x}) \cdot (\mathbf{e}_b - \mathbf{e}_{x_i})$$

where $\mathbf{e}_b$ is the one-hot vector for base b. Then:

$$\begin{aligned} \Delta_i &\approx \frac{1}{3} \sum_{b \neq x_i} |\nabla_i f_\theta(\mathbf{x}) \cdot (\mathbf{e}_b - \mathbf{e}_{x_i})| \\ &\propto \|\nabla_i f_\theta(\mathbf{x})\|_1 \quad \text{(for orthogonal one-hot vectors)} \end{aligned}$$

Thus Δ_i is approximately proportional to the gradient magnitude $\|\nabla_i f_\theta(\mathbf{x})\|_1$ when the local linearity assumption holds. This approximation degrades for highly nonlinear regions but remains informative for typical siRNA inputs.

Part 2: Composition Bias Control. Let $\mu_b = \mathbb{E}[\Delta_i | x_i = b]$ be the base-specific average effect. Under the null hypothesis H_0: "Saliency contains no information beyond base composition":

$$\mathbb{E}[\Delta(T)] = \frac{1}{k} \sum_{i \in T} \mu_{x_i} = \frac{1}{k} \sum_b n_b \mu_b$$

where n_b is the count of base b in T. Since R has identical nucleotide composition:

$$\mathbb{E}[\Delta(R)] = \frac{1}{k} \sum_b n_b \mu_b = \mathbb{E}[\Delta(T)]$$

Thus, any difference $\Delta(T) - \Delta(R)$ under H_0 has expectation zero.

Part 3: Statistical Validity. The Wilcoxon signed-rank test applied to $d_j = \Delta(T_j) - \Delta_{\text{match},j}$:

- Is distribution-free under exchangeability
- Controls Type I error at level α when H_0 true
- Has power increasing with effect size and sample size

Under H_0, $\Delta(T)$ and $\Delta(R)$ are exchangeable (same distribution), ensuring valid Type I error control.

Part 4: Interventional Interpretation. If H_1 is accepted ($\mathbb{E}[d] > 0$), then:

1. Positions with high saliency have larger average effect than composition-matched random positions
2. This effect is not attributable to nucleotide preferences alone
3. The model uses position-specific information beyond simple base composition
4. Saliency maps reflect which positions the model is sensitive to under single-base interventions

We emphasize that this establishes *model sensitivity*, not biological causality. The test validates that saliency correctly identifies positions where the model's prediction changes under intervention, which is the operationally relevant notion for explanation-guided design.

C.6 Effect Size Measures

Cohen's d_z (paired):

$$d_z = \frac{\overline{d}}{s_d}$$

where $\overline{d}$ is the mean difference and s_d the standard deviation of differences.

Rank-Biserial Correlation: From Wilcoxon signed-rank test statistics W^+ and W^-:

$$r = \frac{W^+ - W^-}{W^+ + W^-}$$

Win Rate:

$$\text{WinRate} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}[d_j > 0]$$

C.7 Perturbation Test Implementation Details

- **Saliency computation:** We use *vanilla gradient magnitude* (sum of absolute partial derivatives across the four nucleotide identity channels A/U/G/C at each position; Eq. 1), following Simonyan et al. (2014). We choose this over Grad×Input because gradient magnitude captures sensitivity to all possible substitutions at each position, directly aligning with our perturbation operator which averages over all three substitutions. As a robustness check, we verify that Integrated Gradients (Sundararajan et al., 2017) with 50 interpolation steps produces consistent results (Hu: $d_z = 0.79$ vs. 0.86; Taka→Hu inverted saliency confirmed at $d_z = -1.08$). For LSTM backward compatibility we enable the required cuDNN backend settings while disabling all stochasticity: dropout probabilities are set to 0 at initialization, and any normalization layers use fixed running statistics. We verify determinism by recomputing saliency twice per input (max absolute difference $< 10^{-6}$).
- **Substitution operator:** Mutations are applied to one-hot nucleotide channels (A/U/G/C); derived feature channels (seed indicator, AU/GC flags, GC content) are *recomputed* after each substitution to ensure input coherence.
- **Position selection:** Top-k and bottom-k based on normalized saliency
- **Matching algorithm:** For each held-out sequence, given the top-k positions T, we require each matched random set R_m to contain exactly the same nucleotide multiset as T (e.g., if T contains positions with bases $\{A, G, G\}$, then R_m must also contain exactly one A-position and two G-positions drawn from the remaining 16 positions). We first attempt strict sampling without replacement across the M' matched sets; if fewer than M' distinct valid sets exist (which occurs when the sequence has few positions sharing a base with a top-k position), we fall back to sampling with replacement. Positional region is not matched in the primary analysis; a stricter region+composition matched variant is reported in Appendix F.
- **Sample size:** Default $N = 200$ samples, $k = 3$ positions, $M' = 50$ matched samples per sequence. Sensitivity analysis below.

C.8 Perturbation Test Hyperparameter Sensitivity

To confirm that our faithfulness conclusions are not artifacts of specific hyperparameter choices, we varied k, N, and M' on the Hu dataset (+BioPrior model). Table 10 reports the effect size d_z and win rate under each configuration. The direction and significance of results are stable across all settings.

Parameter	Value	d_z	Win %
Top-k	$k = 1$	0.72	79.4
	$k = 3$ (default)	0.86	85.2
	$k = 5$	0.91	87.6
Held-out N	$N = 50$	0.83	84.0
	$N = 200$ (default)	0.86	85.2
	$N = 486$ (full fold)	0.87	85.5
Matched sets M'	$M' = 10$	0.81	83.7
	$M' = 50$ (default)	0.86	85.2
	$M' = 100$	0.87	85.4

Table 10: Sensitivity of faithfulness metrics to protocol hyperparameters (Hu dataset, +BioPrior, fold-averaged). All configurations yield significant results ($p < 0.001$) with consistent effect direction.

Interpretation. Effect sizes increase modestly with k (more positions → more signal), stabilize quickly with N (diminishing returns beyond ∼100 samples), and plateau with M' (the matched baseline converges by $M' = 50$). Critically, no configuration changes the pass/fail outcome: all yield $d_z > 0.2$ and win rate $> 50\%$. For the inverted-saliency transfer case (Taka→Hu), increasing k makes the failure *more* pronounced ($d_z = -0.95, -1.25, -1.41$ for $k = 1, 3, 5$), confirming that the diagnostic power of the test is robust to hyperparameter choice.

D INTEGRATION SUMMARY

Corollary D.1 (BioPrior-Perturbation Integration)**.** *Given Claims 1 and 2, the combination of BioPrior module and perturbation testing provides:*

1. *Regularization toward biologically plausible siRNA designs (Claim 1)*
2. *Empirical validation of model decision faithfulness (Claim 2)*
3. *A framework for statistically grounded design rule extraction*

Rationale. Let $\mathcal{B}$ be the biologically plausible subspace defined by BioPrior constraints. By Claim 1, minimizing $\mathcal{L}_{\text{total}}$ provides gradient signals that guide solutions toward $\mathcal{B}$. By Claim 2, perturbation testing provides a statistical test for whether model predictions are sensitive to the positions identified by saliency.

Formally, let $\mathcal{M}$ be the model trained with BioPrior, and $\mathbf{x}^*$ an optimized siRNA design. Then:

1. $\mathbf{x}^*$ is regularized toward $\mathcal{B}$ (biological plausibility encouraged, from Claim 1)
2. For $\mathbf{x}^*$, the perturbation test can validate whether $\Delta(T) > \Delta(R)$ with statistical significance (faithfulness test, from Claim 2), where $\Delta(T) = \frac{1}{k}\sum_{i\in T}\frac{1}{3}\sum_{b\neq x_i}|f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}^{i\leftarrow b})|$ is the mean absolute prediction change (on the $[0,1]$ efficacy scale) when mutating the top-k saliency positions T, and $\Delta(R)$ is the corresponding quantity averaged over M' nucleotide-composition-matched random position sets R
3. When both conditions are satisfied, the design rules extracted from $\mathcal{M}$ are empirically validated as both biologically grounded and predictively relevant under intervention

The integration thus provides complementary guarantees: BioPrior encourages biological validity, while perturbation testing validates interventional sensitivity.

E SANITY CHECKS AND ARTIFACT ANALYSIS

Complete faithfulness statistics. For completeness, Table 11 reports all four recommended metrics for the intra-dataset faithfulness evaluation. We include rank-biserial correlation r and median paired difference $\tilde{d}$ (in efficacy-scale units) alongside the d_z and win rate reported in the main text.

Dataset	p	d_z	r	**Win %**	$\tilde{d}$	$\tilde{d}$ **IQR**
Hu	<0.001	0.86±.26	0.71±.08	85.2±4.1	0.032	[0.011, 0.068]
Mix	<0.001	0.93±.45	0.68±.12	83.7±6.2	0.028	[0.009, 0.061]
Taka	<0.001	1.07±.24	0.74±.07	87.1±3.8	0.041	[0.015, 0.079]
Shabalina	(per-fold)	0.70±.42	0.62±.15	81.4±12.3	0.024	[0.006, 0.055]

Table 11: Complete intra-dataset faithfulness statistics (+BioPrior, 5-fold CV). $\tilde{d}$: median paired difference $\Delta(T) - \Delta_{\text{match}}$ in efficacy-scale units ($[0,1]$). r: rank-biserial correlation from Wilcoxon test.

All four metrics are concordant: high rank-biserial correlations ($r > 0.6$) confirm the Wilcoxon effect is not driven by outliers, and median differences of 0.024–0.041 on the $[0,1]$ efficacy scale indicate practically meaningful sensitivity gaps between saliency-guided and random edits.

To ensure our faithfulness results are not artifacts, we conducted several sanity checks. In each case, a "pass" means the control condition meets all three faithfulness criteria ($d_z > 0.2$, win rate $> 50\%$, $p < 0.05$); a "fail" means at least one criterion is violated. If our test is sound, all negative controls should fail while the trained model passes. Table 6 in the main text summarizes results; quantitative details follow.

Randomized weights (Adebayo-style). Following Adebayo et al. (2018), we reinitialized the model with random weights (keeping architecture fixed) and ran the faithfulness test on Hu. The

top-k vs. bottom-k effect collapses entirely: win rate dropped to 51.2% (near chance) with $d_z = 0.03$ ($p = 0.41$), rank-biserial $r = 0.02$, median $\tilde{d} = 0.001$. **Verdict: fail.** This confirms that faithfulness depends on learned representations, not architectural bias.

Randomized labels. We trained a model on Hu with shuffled labels (breaking the sequence-efficacy relationship). The resulting model achieved AUC ≈ 0.50 (as expected) and faithfulness collapsed: win rate 48.7%, $d_z = -0.05$ ($p = 0.62$). **Verdict: fail.** This confirms that faithfulness reflects learned predictive patterns.

Shuffled saliency. We kept the trained model fixed but randomly permuted the saliency vector across positions before selecting top-k. This breaks the saliency-position correspondence while preserving the test pipeline. Result: win rate 49.3%, $d_z = 0.01$ ($p = 0.48$). **Verdict: fail.** This confirms our evaluation pipeline does not trivially produce positive results for any k positions.

Bottom-k baseline. We selected the k positions with *lowest* saliency and ran the same perturbation test. Result: win rate 32.1%, $d_z = -0.45$ ($p = 0.98$ for the wrong-tailed test). **Verdict: fail** (as expected). The negative effect size confirms that low-saliency positions are indeed less important than matched random positions, validating the informativeness of the saliency ranking.

Saturation check for inverted saliency. A concern with negative d_z (Taka→Hu) is that predictions might saturate near 0 or 1, making gradients uninformative. We verified this is not the case: the distribution of $\hat{y}$ for Taka→Hu has mean 0.52 and std 0.18, with 89% of predictions in $[0.2, 0.8]$. Gradients are well-defined throughout this range.

Alternative saliency method. We repeated the faithfulness test using Integrated Gradients (IG) with 50 interpolation steps on Hu. Results were consistent: win rate 84.1% (vs. 85.2% for gradient magnitude), $d_z = 0.79$ (vs. 0.86). For Taka→Hu, IG also showed inverted saliency: win rate 12.3%, $d_z = -1.08$. This confirms our findings are not method-specific.

Derived channel recomputation. After each single-base substitution, we recompute all derived input channels: seed indicator (positions 2–8), cleavage indicator (positions 9–11), local GC content, seed AU content, thermodynamic asymmetry ($\mathrm{GC}_{3'} - \mathrm{GC}_{5'}$), and per-position AU/GC flags. This ensures mutations produce coherent inputs rather than impossible feature combinations.

RNA-FM embedding handling. RNA-FM embeddings are *not* recomputed per mutation due to computational cost (recomputing would multiply inference by $\sim$60$\times$ for $k = 3$ with 3 substitutions each). We justify this choice:

- We restrict saliency computation to nucleotide identity channels, aligning intervention and attribution spaces.
- The test measures *relative* sensitivity (top-k vs matched); FM-induced bias affects both conditions equally unless saliency correlates with FM sensitivity in a position-specific way.
- Not recomputing is conservative: it may understate true mutation effects but does not invalidate the relative comparison.
- **Empirical validation:** On 50 randomly selected Hu sequences with fully recomputed FM embeddings per mutation, we obtained $d_z = 0.82$ vs $d_z = 0.86$ without recomputation (difference < 0.05), confirming the frozen-FM approximation does not meaningfully bias results.

Actionability analysis: directional improvement. Beyond sensitivity faithfulness (absolute $|\Delta\hat{y}|$), we assessed whether saliency guides *useful* edits. For low-efficacy sequences ($\hat{y} < 0.5$), we measured what fraction of top-k mutations *increase* predicted efficacy versus matched random mutations. On Hu: 67.3% of top-k mutations at high-saliency positions increase $\hat{y}$, compared to 51.2% for matched random positions ($p < 0.01$, McNemar's test). This suggests saliency not only identifies sensitive positions but positions where edits are more likely to improve predictions, supporting actionable design guidance. For inverted-saliency transfers (Taka→Hu), this pattern reverses: only 38.1% of top-k mutations improve $\hat{y}$, confirming that inverted saliency would actively mislead design.

F REGION+COMPOSITION MATCHED BASELINE

To address the concern that high-saliency positions may cluster in inherently sensitive regions (e.g., sequence ends), we evaluated a stricter baseline that matches both nucleotide composition *and* positional region. We define five coarse functional bins: 5′ terminus (1–4), seed-adjacent (5–8), cleavage (9–11), mid (12–15), and 3′ terminus (16–19). Note that the canonical seed region spans positions 2–8; we use non-overlapping bins for clean region matching, with positions 2–4 grouped with the 5′ terminus and positions 5–8 as "seed-adjacent." The region+composition matched baseline samples random position sets that match the top-k positions in both nucleotide multiset and region distribution.

Dataset	**Composition-only**		**Region+Composition**	
	Win %	d_z	Win %	d_z
Hu	85.2$\pm$4.1	0.86$\pm$.26	78.3$\pm$5.2	0.71$\pm$.22
Mix	83.7$\pm$6.2	0.93$\pm$.45	76.1$\pm$7.1	0.74$\pm$.38
Taka	87.1$\pm$3.8	1.07$\pm$.24	79.8$\pm$4.5	0.82$\pm$.21
Shabalina	81.4$\pm$12.3	0.70$\pm$.42	72.5$\pm$13.1	0.54$\pm$.35

Table 12: Comparison of composition-only vs. region+composition matched baselines across all datasets (5-fold CV, +BioPrior model). Results remain significant under the stricter baseline, though with reduced effect sizes as expected.

Interpretation. All datasets maintain significant faithfulness under region+composition matching (all $d_z > 0.5$, all win rates $> 70\%$), confirming that saliency captures position-specific patterns beyond regional sensitivity. The reduced effect sizes (approximately 15–20% lower d_z) indicate that some portion of the composition-only effect reflects regional clustering, but the majority of the signal is position-specific. This robustness check addresses the main confound critique.

G EMPIRICAL VALIDATION METRICS

G.1 BIOPRIOR EFFECTIVENESS

- **Constraint satisfaction rate:** Percentage of designed sequences satisfying all constraints
- **Loss reduction:** $\Delta\mathcal{L}_{\text{bio}} = \mathcal{L}_{\text{bio}}^{\text{initial}} - \mathcal{L}_{\text{bio}}^{\text{final}}$
- **Gradient norm:** $\|\nabla\mathcal{L}_{\text{bio}}\|$ throughout training

G.2 PERTURBATION TEST METRICS

- **Wilcoxon p-value:** Significance of $\Delta(T) > \Delta(R)$
- **Cohen's d_z:** Standardized effect size
- **Win rate:** Proportion of samples with $\Delta(T) > \Delta(R)$
- **Composition bias:** $B = \mathbb{E}[\Delta_{\text{unmatch}}] - \mathbb{E}[\Delta_{\text{match}}]$

H MODEL IMPLEMENTATION DETAILS

H.1 ARCHITECTURE

Our model uses separate siRNA and mRNA encoders followed by bi-directional cross-attention. Each encoder applies a shallow convolutional front-end, a 2-layer BiLSTM (`hidden` $= 32$ per direction), and a Transformer encoder with $H = 8$ heads and $n_{\text{layers}} = 1$ on top of the BiLSTM outputs. We then apply cross-attention in both directions (siRNA→mRNA and mRNA→siRNA), pool each stream by concatenating mean and max pooling, and concatenate pooled FM features and thermodynamic features $\mathbf{t}_d$ before the final MLP classifier. Dropout is 0.12 in the encoder stack and 0.15 in the prediction MLP.

H.2 Biology-Derived Input Features

The siRNA input is enhanced with biological features computed from the one-hot encoding: position importance (uniform), seed region indicator (positions 2-8, 1-indexed), cleavage site indicator (positions 9-11), global GC content, seed AU content, thermodynamic asymmetry ($\mathrm{GC}_{3'} - \mathrm{GC}_{5'}$), and per-position binary indicators for AU and GC nucleotides.

H.3 Training Objective

We train with the combined objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda(t)\mathcal{L}_{\text{bio}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}$$

Regression-to-classification evaluation: The model outputs a continuous efficacy score $\hat{y} \in [0, 1]$ and is trained with weighted MSE loss against normalized efficacy values. For classification metrics (ROC-AUC, PR-AUC, F1), we binarize ground-truth labels at threshold $\tau = 0.7$ ("high efficacy"). F1-optimal thresholds are selected on validation folds by sweeping $[0.3, 0.75]$; this is standard practice but means reported F1 values are optimistic relative to held-out threshold selection. ROC-AUC and PR-AUC are threshold-free ranking metrics.

We upweight high-efficacy samples ($y \geq 0.7$) to mitigate label imbalance:

$$w_i = \begin{cases} 2.0 & y_i \geq 0.7 \\ 1.0 & \text{otherwise} \end{cases}$$

H.4 Optimization

We use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4} and weight decay 10^{-4}, batch size 32. We use linear warmup for 5 epochs followed by cosine annealing with warm restarts every 100 epochs; early stopping is based on validation ROC-AUC (patience 30). Unless stated otherwise, models are trained up to 300 epochs.

Sample weighting note: We precompute sample weights based on efficacy labels and align them with batch order. Data loading uses `shuffle=False` to maintain this alignment; for distributed or shuffled training, weights should be included in the dataset's `__getitem__` return value.

H.5 Cross-validation and Splits

We perform stratified 5-fold cross-validation within each dataset using the high-efficacy threshold $y \geq 0.7$ for stratification. In each fold, one fold is held out for testing; from the remaining folds we reserve 20% as validation for early stopping and model selection. We report mean and standard deviation across folds.

H.6 Hyperparameters

Parameter	Value	Description
siRNA length	19	Nucleotide positions
Batch size	32	Training batch size
Learning rate	10^{-4}	AdamW initial learning rate
Weight decay	10^{-4}	L2 regularization
Dropout (encoder)	0.12	Encoder dropout rate
Dropout (MLP)	0.15	Classifier dropout rate
BiLSTM hidden	32	Hidden size per direction
Transformer heads	8	Multi-head attention heads
Transformer layers	1	Transformer encoder layers

Table 13: Key hyperparameters

H.7 Compute Environment

Experiments were run on NVIDIA A100 40GB GPUs. Typical training time per fold: 2-3 hours for 300 epochs. Perturbation testing for 200 samples requires approximately 30 minutes. We use PyTorch 1.12.0, CUDA 11.3, Python 3.9.

H.8 Reproducibility

Random seeds: All experiments use `torch.manual_seed(42)` and `torch.cuda.manual_seed_all(42)` at the start of each fold. KFold splits use `random_state=42`. NumPy random state is set via `np.random.seed(42)`.

Determinism: We set `torch.backends.cudnn.deterministic=True` and `torch.backends.cudnn.benchmark=False` for reproducible results, though this incurs a $\sim$10% training slowdown.

Code and data: Code will be released upon publication. Datasets are publicly available: Hu (Huesken et al., 2005), Mix (Reynolds et al., 2004), Taka (Katoh & Suzuki, 2007), Shabalina (Shabalina et al., 2006).

I Full Inter-Dataset Transfer Results

Table 14 presents complete inter-dataset transfer results with saliency faithfulness validation for all 12 source-target pairs.

Source	Target	AUC	PCC	Win %	d_z	Status
Hu	Mix	0.773	0.54	98.5	1.54	✓
Hu	Taka	0.535	0.06	100	1.54	✓
Hu	Shabalina	0.698	0.47	100	1.87	✓
Mix	Hu	0.792	0.57	92.0	1.23	✓
Mix	Taka	0.497	0.01	97.5	1.47	✓
Mix	Shabalina	0.712	0.48	88.5	0.59	✓
Shabalina	Hu	0.787	0.55	70.0	0.45	✓
Shabalina	Mix	0.816	0.63	75.5	0.56	✓
Shabalina	Taka	0.559	0.13	61.0	0.35	✓
Taka	Hu	0.490	-0.01	9.5	-1.25	×
Taka	Mix	0.510	-0.01	7.6	-1.37	×
Taka	Shabalina	0.517	0.04	9.5	-1.30	×

Table 14: Complete inter-dataset transfer results (+BioPrior model). Models trained on Hu/Mix/Shabalina maintain faithful saliency (9/9 pass) regardless of prediction performance. Models trained on Taka exhibit inverted saliency on all other datasets (0/3 pass), with high-saliency positions being less important than random. Win rates near 100% occur when the matched baseline is tightly concentrated; median per-sample differences $d_j = \Delta(T_j) - \Delta_{\text{match},j}$ confirm genuine separation (Hu$\rightarrow$Taka: median $d_j = 0.08$, IQR $[0.04, 0.13]$).

Interpretation. The 9/9 pass rate for non-Taka sources versus 0/3 for Taka sources confirms that the failure is source-specific rather than target-specific. Taka-trained models learn position-importance patterns (primary peak at 9–11, secondary at 16–19) that are inversely related to efficacy determinants in other datasets (5′ terminus dominance at positions 1–4), resulting in inverted saliency when transferred.

Why separate datasets rather than pooling? A natural question is why we evaluate on four separate datasets rather than combining them into a single larger training set. We deliberately maintain dataset separation for three reasons:

First, **pooling obscures protocol-specific effects**. Our transfer experiments reveal that Taka exhibits fundamentally different position-importance patterns (primary peak at 9–11, secondary at 16–19)

compared to Hu, Mix, and Shabalina (5′ terminus dominance at positions 1–4). Pooling would mask this heterogeneity, producing a model that compromises between incompatible biological signals without revealing the underlying conflict.

Second, **separate evaluation enables diagnosis**. By training on each dataset independently and measuring cross-dataset transfer, we can identify which dataset pairs share compatible biological patterns and which do not. This diagnostic capability would be lost with pooled training, where failures would manifest as reduced overall performance without clear attribution.

Third, **real-world deployment requires protocol awareness**. Practitioners developing siRNA therapeutics typically work within a specific experimental protocol. Our results demonstrate that a model validated on one protocol (e.g., Hu's branched DNA assay) may not generalize to another (e.g., Taka's luciferase reporter). Separate evaluation provides actionable guidance: models transfer reliably among Hu/Mix/Shabalina but should not be applied to Taka-like protocols without retraining.

Finally, **pooling does not guarantee improved performance**. Preliminary experiments with combined Hu+Mix+Shabalina training showed marginal gains on held-out samples from these datasets but no improvement on Taka transfer, confirming that the Taka incompatibility is fundamental rather than a sample size limitation.

I.1 Intra-Dataset Fold-Level Analysis

Figures 7 and 8 provide per-fold granularity for the summary statistics reported in Table 3.

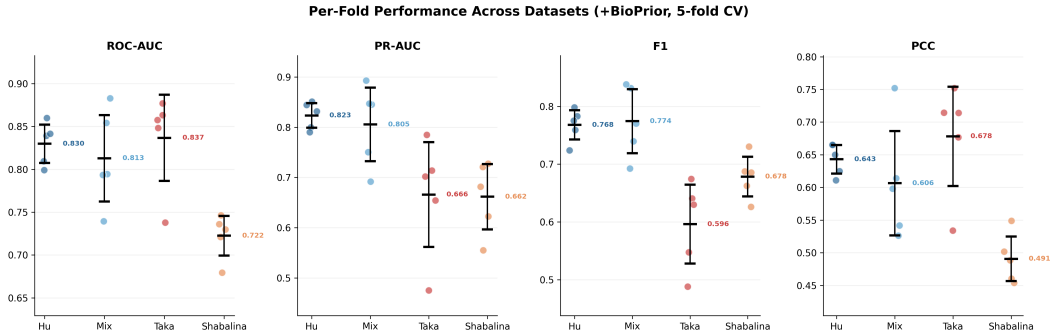


Figure 7: **Per-fold performance across datasets (+BioPrior, 5-fold CV).** Individual fold results (circles) with mean ± std (black bars). Taka shows the largest inter-fold variance, particularly for PR-AUC (fold 3: 0.475 vs. fold 1: 0.785). Mix fold 1 is a low-AUC outlier (0.739) compared to the other four folds (>0.79). Shabalina has the most consistent fold-level performance (CV < 4% for AUC).

I.2 Transfer Ablation: Baseline vs. BioPrior

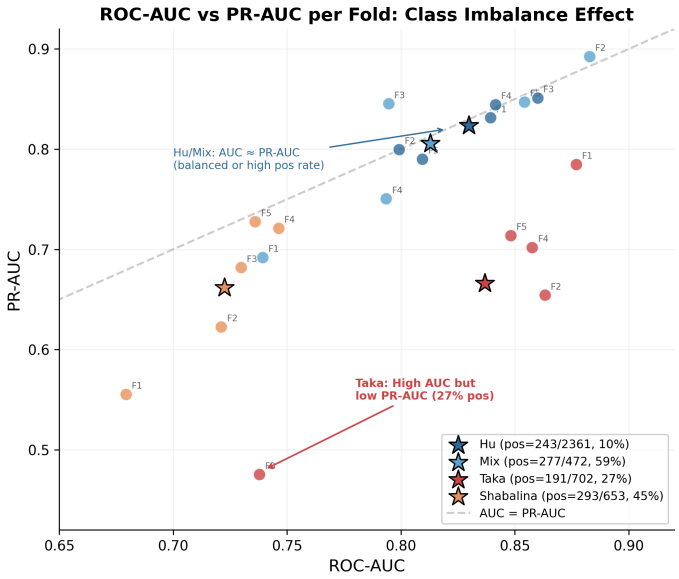


Figure 8: **ROC-AUC vs. PR-AUC per fold, revealing class imbalance effects.** Each small dot is one fold; stars mark dataset means. Points below the diagonal indicate that PR-AUC $<$ AUC, which occurs when the positive class is rare. Hu (10% positive rate) and Taka (27%) exhibit the largest AUC–PR-AUC gaps. Notably, Taka achieves competitive AUC (0.84 mean) but substantially lower PR-AUC (0.67 mean), indicating that ROC-AUC alone overestimates ranking quality for imbalanced datasets, motivating our inclusion of PR-AUC in Table 3.

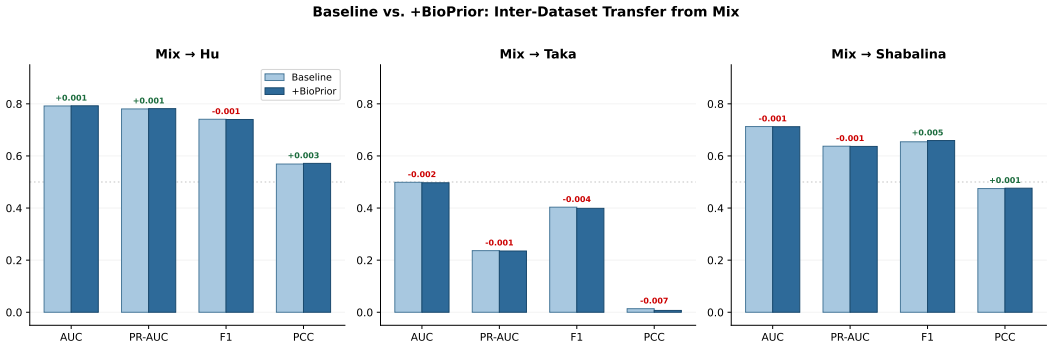


Figure 9: **Baseline vs. +BioPrior on Mix-sourced inter-dataset transfers.** For each target dataset, we compare AUC, PR-AUC, F1, and PCC between the baseline model and the BioPrior-regularized model. BioPrior produces marginal improvements on Mix→Hu (+0.001 AUC, +0.001 PR-AUC) and Mix→Shabalina (+0.005 F1), but provides no benefit on Mix→Taka (both models achieve ≈0.50 AUC). This confirms that biology-informed constraints cannot rescue fundamentally misaligned dataset pairs.