DECAY SIGNATURES: MEASURING DOMAIN-DEPENDENT DEGRADATION AND THE BOUNDARIES OF ANCHOR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training on degraded inputs is a standard route to robust representations, yet choosing the degradation intensity remains empirical: existing approaches scan, search, or offer qualitative principles — none compute a schedule from measurable properties of its operators. Working in cross-modal molecule image–graph retrieval, where structure is discrete and verifiable, we introduce *decay signatures*: per-operator decay rates measured on task, structure, and frequency channels. Signatures overturn conventional operator labels — operator class is a measured, domain-dependent fact, not a priori intuition — and reveal a non-monotone, U-shaped response of fixed-anchor training to intensity, which decomposes into a mild-anchor penalty and a deployment-conditional optimum. Stress-testing a first-order transition-landmark predictor (product of per-operator survival curves, threshold calibrated once) on ten held-out schedules under pre-registration yields a boundary map: the primary criterion is not met (2/4 under the plateau reading, 1/4 under strict argmax scoring); prediction is exact at bracket resolution on a restricted subset, with the caveat that one exact hit rests on a near-cancellation of two measured error terms; and failures decompose into three observed structural sources — saturation, interaction error, and compensability. Degradation intensity is measurable — but only to first order; we report this boundary with the same fidelity as the hits.

1 INTRODUCTION

Training with corrupted, degraded, or augmented inputs is one of the oldest and most effective routes to robust representations. Yet a basic design question remains surprisingly unprincipled: *how strong should the degradation be*? Existing answers fall into three families — empirical intensity sweeps, learned or searched augmentation policies (Yang et al., 2025; 2024; Cubuk et al., 2019; Wang et al., 2021), and qualitative information principles (Tian et al., 2020) — and all three must be redone for every new operator set, every new task, and every new coupling; none computes a schedule from measurable properties of its operators.

This paper asks whether the anchor — the training degradation intensity — can instead be *computed from measurements of the operators themselves*. The starting point is an empirical observation in cross-modal molecular image–graph retrieval: under asymmetric training, where degraded images are trained against clean graph targets, fixed-anchor training yields a non-monotone, U-shaped response to intensity — qualitatively anticipated by sweet-spot principles (Tian et al., 2020), but with landmarks no existing account locates — with a reproducible valley and recovery peak across five seeds (Figure 1). The decision-relevant layer is larger: at fixed deployment intensity, self-matched near-optimal anchors lose 4–31 R@1 points, and the optimal anchor shifts harsher as deployment hardens. Locating this response's landmarks in advance requires knowing how fast each information channel of the input decays — which no existing instrument provides.

We therefore build the account bottom-up, measurement first. **Decay signatures** (Section 4) record, for each operator, how fast task performance, structure, and frequency content decay as intensity increases. Measurement already delivers one surprise before any prediction is attempted: operator classes turn out to be measured, domain-dependent facts, not a priori labels — blur, conventionally

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

a texture purifier, damages the high-frequency foreground structure of molecular line drawings, and on CIFAR-10 the classification flips, exactly as the frequency argument predicts. The same measurement rules out smooth rate models: destruction is threshold-like, and exponential decay families underfit at $R^2 \leq 0.28$ — a negative result we report.

We then stress-test the simplest predictor consistent with what the measurement rules out — composite structure-survival as the product of per-operator curves, with a threshold calibrated once — under pre-registration on ten held-out schedules, and report the resulting boundary map at the same fidelity as its hits: the pre-registered primary criterion is not met (2/4 under the plateau reading, 1/4 under strict argmax scoring); prediction is exact at bracket resolution on a restricted subset; and failures decompose into three observed structural sources — saturation, interaction error, and compensability (Section 5).

The paper delivers three things: a measurement instrument (Section 4), a characterization of the phenomenon that motivates it (Section 3), and a precise map of where a first-order estimator built on these measurements succeeds, fails, and why (Section 5). For practitioners the deliverable is an explicit three-regime decision rule (§5.4).

2 Related Work

Augmentation and corruption theory. Kernel and group-theoretic accounts explain when augmentation reduces variance without adding bias (Dao et al., 2019; Chen et al., 2020); feature-level analyses show augmentation acts asymmetrically on feature types, suppressing redundant or misleading features while sparing semantic ones (Wen & Li, 2021; Shen et al., 2022; von Kügelgen et al., 2021). Closest to our measurement layer, Lin et al. (2024) prove that augmentation induces implicit *spectral* operations on the data covariance that can be good, bad, or ugly depending on the data spectrum — a theoretical counterpart of our finding that operator classes depend on the domain. The differences are load-bearing: their spectrum is the covariance eigenspectrum of linear models, while our decay signatures are empirical multi-channel measurements in a cross-modal contrastive setting; they predict no intensity, and we locate the transition landmarks of the bias–variance tradeoff they characterize. Kirichenko et al. (2023) show augmentation harms are class-asymmetric due to distribution overlap and must be fixed at the *source* class — independent support for our source-side (image-tower) intervention with a clean target anchor. Two results bound what can be claimed from structure-based explanations: perceptual similarity between training augmentations and test corruptions (Hendrycks & Dietterich, 2019) predicts corruption error and fails on dissimilar corruptions (Mintun et al., 2021), and increased shape bias does not by itself cause corruption robustness (Mummadi et al., 2021). Our claims are compatible with both: we predict intensity within one coordinate system (Section 3.1) and measure structure preservation directly (ring-match on ground truth) rather than claiming a bias-direction mechanism. In the molecular domain, graph contrastive augmentations already require domain-aware design (You et al., 2020; Wang et al., 2022), and motif-aware pretraining shows that generic contrastive augmentations can destroy motif information (Liu et al., 2025), reinforcing our clean-anchor design.

Choosing the intensity: search vs. prediction. All existing routes to intensity selection are empirical or learned: policy search (Cubuk et al., 2019), reduced-search-space sampling (Cubuk et al., 2020; Müller & Hutter, 2021; Hendrycks et al., 2020), learned corruption generators (Hendrycks et al., 2021), learned magnitude schedules over training (Ho et al., 2019) and single-knob annealing curricula (Sinha & Garg, 2020), adversarial mixing (Wang et al., 2021), adversarial noise training (Rusak et al., 2022), information-theoretic view search (Tian et al., 2020), entropy-driven (Yang et al., 2024) and RL-driven (Yang et al., 2025) per-sample adaptation, weight-space perturbations that mimic input corruptions (Trinh et al., 2024), and distribution-level interventions that rebalance or protect informative content (Nguyen et al., 2024; Yang et al., 2026). Measurement-based criteria quantify augmentation properties without localizing intensity optima (Gontijo-Lopes et al., 2020). A related line of work closes the loop between measurement and intensity adaptively: RangeAugment learns per-operation magnitude ranges from an image-similarity signal (Mehta et al., 2023), Suzuki (2022) steers augmentation search toward transformations a teacher model finds informative, and adaptive discriminator augmentation tunes the augmentation probability online from a measured discriminator overfitting heuristic (Karras et al., 2020). These methods are measurement-driven, but the measurement is retaken online during each training run and the output is a per-sample or per-

run adaptation rather than an explicit schedule. Our aim is complementary: we ask whether the landmarks of a training schedule can be read off from one-time operator measurements taken before training, and whether that read-out transfers to held-out composite schedules. To our knowledge, no verified *pre-computation* of the optimal training intensity from measured operator properties has been demonstrated on held-out schedules: these methods either learn the policy online, reduce the problem to one or two strength knobs found by search, or remain qualitative principles. Two recent theoretical results come closest to computing augmentation quantities from measurable properties: an augmentation-aware error bound whose two distance terms predict optimal SSL augmentation parameters (Cui et al., 2025), and a group-diameter control parameter that tracks the generalization gap (Bouyahia et al., 2026) — both are per-task criteria evaluated on the training distribution, and neither is verified as a predictor on held-out schedules. Our predictor is the complement: measurement amortizes coarse landmark localization within a fixed coordinate system, with a pre-registered verification protocol.

Degradation estimation. Estimating degradation parameters from a degraded input is a well-studied problem in image restoration. Blind super-resolution methods estimate the downsampling kernel from internal statistics of a single image (Bell-Kligler et al., 2019) or synthesize realistic degradation pipelines to train restoration networks (Zhang et al., 2021), and classical estimators recover noise levels (Liu et al., 2006; 2013) or JPEG compression quality (Wang et al., 2002) from a single image without reference. In this literature, measurement serves restoration: the estimated parameters drive the reconstruction of the degraded input. We use operator-level measurement for a different purpose — not to undo the degradation on a given input, but to predict where, along an intensity axis, structural information stops being learnable during training. The two uses are complementary.

Non-monotonic learning dynamics. Double descent (Nakkiran et al., 2020) and grokking (Power et al., 2022; Liu et al., 2023) established non-monotonic generalization as a real phenomenon class, unified by two-speed learning accounts (Davies et al., 2023; Huang et al., 2024); robustness itself can emerge late and non-monotonically (Humayun et al., 2024). We add a new axis — input degradation intensity — and, unlike this family, a pre-registered predictor with an explicit, measured boundary map rather than a post-hoc account.

Shortcut learning. Models preferentially learn simple, spurious features (Shah et al., 2020; Geirhos et al., 2020) because dominant features starve others of gradient (Pezeshki et al., 2021), and strong spurious correlations are not forgotten once learned (Qiu et al., 2024; Bombari & Mondelli, 2025) — together implying that input-side removal is the effective intervention point, which is where our degradation acts. Izmailov et al. (2022) show ERM representations already contain core features; our clean cross-modal anchor can be read as a ground-truth version of that readout fix, transposed to extremely degraded inputs. Degradation training compresses target-irrelevant information on the input side, complementary to latent-space compression objectives (Kawaguchi et al., 2023; Huang et al., 2026).

3 Setup and the Non-Monotonic Robustness Curve

3.1 Setting

We study cross-modal molecular retrieval on 200,000 PubChem image–graph pairs (split 193k/2k/5k): a CLIP-style dual-tower contrastive network (Radford et al., 2021; Oord et al., 2018) with a ConvNeXt-Base image tower (Liu et al., 2022) and a GIN graph tower (Xu et al., 2019) over FastRP embeddings (Chen et al., 2019), trained *asymmetrically* — the image tower sees degraded inputs while the graph tower always sees clean graphs; this asymmetry is load-bearing, and we return to it in Section 6. Degradation is parameterized by a continuous intensity L driving a composite of four physical operators — blur, JPEG, downscale, and coarse dropout — through a fixed knob mapping; a coupling parameter scales purifier-side and destructor-side knobs asymmetrically (Section 5). L is an *operational coordinate*; its absolute scale is a display choice, and all claims are made and tested within this coordinate system. We report two protocols: the self-matched diagnostic (test L = training L) and the primary fixed-deployment view (fixed test L). Unless stated otherwise,

numbers are over 5 seeds. The nine degradation operators are characterized by measurement in Section 4; all protocol details are in the appendix.

Scope. Our claims concern *intensity selection within this coordinate system*, not robustness to unseen corruption types. Perceptual similarity between training and test corruptions predicts corruption error but fails on dissimilar corruptions (Mintun et al., 2021); our held-out schedules are precisely those cases, so similarity-based explanations cannot locate the optimum.

A one-minute domain primer. Molecular structure drawings are sparse line art: bond lines and ring edges are thin dark strokes on blank background — high-frequency content. The task metric is image-to-graph retrieval R@1, and the structure probe, *ring-match*, is the fraction of test molecules whose ring count is exactly recovered by the top-1 retrieved graph. Ring-match is computed through the retrieval model — a task-level behavioral measurement, not an operator-intrinsic constant (Appendix C).

3.2 A NON-MONOTONE RESPONSE TO DEGRADATION INTENSITY

Fixed-anchor training — the *anchor* throughout denotes the fixed training degradation intensity, not the query view of the contrastive loss; the clean graph target plays the latter role here — responds non-monotonically to degradation intensity (Figure 1a). Under the self-matched protocol, the composite schedule holds a high plateau through moderate intensities (plateau maximum 93.88 at L=0.124), dips at $L_c \approx 0.14$ (mean minimum 93.06 ± 0.53 at L=0.138), and recovers to a post-dip local peak at $L_{\text{opt}} = 0.152$ (93.20 ± 0.72), before declining at extreme intensity. The dip-and-recovery is shallow in magnitude but consistent in every seed: per-seed minima fall in $L \in [0.140, 0.146]$, and every seed recovers from its minimum to a post-dip peak in $L \in [0.148, 0.152]$, gaining $+0.40$ to $+1.64$ R@1 points (mean $+0.93$; paired $t(4) = 4.59$, $p = 0.010$; Wilcoxon $p = 0.0625$ at n=5, the attainable floor, so we report per-seed effect sizes). Beyond the plateau edge, the decline is abrupt and partly stochastic: $\text{std} = 4.2$ at L=0.160 (driven by a single collapsed seed, 83.0 vs. 91.7–92.8 for the other four) against ≤ 0.8 across the plateau (Section 5.4). We stress the protocol: this self-matched view is a *training-dynamics diagnostic* — it conflates anchor quality with test difficulty.

The deployment view (Figure 1b) holds the test intensity fixed and is our primary evidence. It resolves the response into two layers. *First, a mild-anchor penalty*: at test L=0.150, anchors below 0.10 lose 4–31 R@1 points relative to the self-matched optimum ($0.110 \rightarrow 89.44 \pm 1.94$ vs. 93.35 at L=0.152; paired $t(4) = 3.98$, $p = 0.016$) — anchors that look nearly optimal self-matched are far from optimal at extreme deployment. *Second, a deployment-conditional optimum*: at test L=0.150 the response forms a broad plateau spanning $[0.13, 0.20]$ (peak 93.44 ± 0.18 at 0.162), while at test L=0.200 it increases throughout the scanned range, so the best anchor is the harshest trained (77.86 ± 0.95 at 0.200). Together they rule out both "deeper is better" and "milder is better", and recast anchor selection as a per-deployment decision — which motivates testing whether such a decision can be computed from operator measurements at all (Section 5). Provenance: Figure 1b uses the fine-scan pool (test 0.150); the dedicated controls (Appendix Table 7) are an earlier vintage (test 0.152), hence the $\sim$2-point offset; within-panel comparisons are same-protocol throughout.

3.3 THE PHENOMENON IS SPECIFIC: CONSTRAINTS ON ANY EXPLANATION

Controls (appendix) show the response is specific and constrain what any explanation must account for: (i) a *mixture* baseline (MixTrain, uniform random intensity per sample) trained on the same operators shows a *monotone* response — no valley — so the U-shape is a property of fixed-anchor training dynamics, not of corruption exposure per se; (ii) a frequency-matched control, applying the same low-pass operators at randomized intensities, collapses — the valley is not low-pass filtering; (iii) on a music retrieval task where texture (timbre) and structure (pitch, rhythm) are decoupled by construction, no U-shape is observed (anecdotal, qualitative only). One reading is channel competition: texture and structure compete for representational capacity during training (Pezeshki et al., 2021), and the valley marks the region where the texture channel wins. We do not validate this mechanism. Instead we extract its measurable consequence — that per-channel decay rates are the quantities a predictive theory should be built from — which leads directly to the measurement instrument of Section 4.

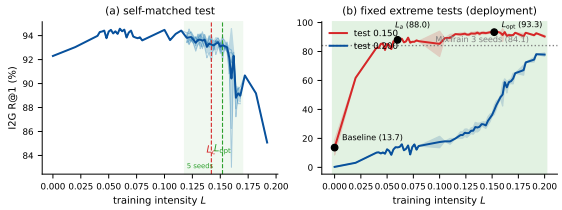

Figure 1: **Non-monotonic robustness under increasing training intensity.** (a) Self-matched evaluation (test L = training L), a training-dynamics diagnostic: valley at $L_c \approx 0.14$, post-dip local peak at $L_{\text{opt}} = 0.152$; the dip at L=0.160 is run variance (one of five seeds collapses). Per-seed values: supplementary `fig1_valley_perseed.csv`. (b) Primary deployment view at fixed test L=0.150 and L=0.200 (multi-seed means, 95% CI bands): mild anchors that look preferable self-matched collapse under extreme deployment; the optimal anchor shifts harsher as deployment hardens. Shaded window: multi-seed fine scan; coarse points outside are single-run.

4 Decay Signatures: Measuring Degradation Operators

The constraints of Section 3.3 point to one kind of quantity: per-channel decay rates. This section defines the instrument that measures them.

4.1 The instrument

For each operator T and intensity L, we apply the single-operator degradation $T(\cdot; L)$ to the test split and record three channels on a fixed reference model trained on clean, undegraded data:

- **Task channel** (α_{task}): decay of retrieval R@1 of the clean-trained reference model as L increases.
- **Structure channel** (α_{topo}): decay of task-level structure preservation under a fixed reference model, measured as the fraction of molecules whose ring count is exactly recovered by the retrieved graph — the ring-match rate. Ring-match verifies ring topology only — not atom types, bond orders, or stereochemistry — and is a first-order structural probe (Section 6).
- **Frequency channel**: change of the high-frequency energy ratio of the degraded image, computed over the outer half of the FFT radial spectrum — a model-free image statistic (Yin et al., 2019) that indicates whether an operator *removes* or *injects* high-frequency content.

Two protocol choices are load-bearing. First, rates are estimated on the *operating window* $L \in [0.02, 0.15]$, the regime in which composite schedules operate, using a functional-form-free two-point log-rate $\alpha = \ln(y_{L_1}/y_{L_2})/(L_2 - L_1)$, since threshold-type morphologies make parametric fits unreliable (Section 4.3). Second, all measurements use the same reference checkpoint, so signatures characterize the operator, not a trained robustness response; reference-model sensitivity is quantified in the appendix. Knob effective intensities and saturation semantics for coupled schedules are defined in Section 5 and the appendix.

4.2 Operator class is measured, not assumed

Applied to the nine operators, the signatures overturn the conventional taxonomy (Figure 2, Table 1). The developed example is blur. Conventional usage treats blur as a texture purifier (Geirhos et al., 2019): it removes high-frequency noise and little else. In this domain, however, the foreground structure of a molecular line drawing — bond lines, ring edges — is itself high-frequency, and the structure channel confirms the damage: blur's ring-match decay rate ($\alpha_{\text{topo}} = 2.22$) is of the destructors' order, an order of magnitude above any purifier's. Blur is a purifier only in domains whose foreground is low-frequency. Of the nine operators, exactly one — JPEG — qualifies as a clean purifier in this domain, with a gentle task-channel rate (0.71) and a nearly flat structure channel (0.63). Additive-noise and occlusion operators are *maskers*: they preserve structure in the operating window but inject high-frequency content. All labels follow a rule fixed before any downstream use

Table 1: Measured operator signatures on the molecular task (operating window). α values are two-point log-rates over exact measured endpoints (L=0.02, 0.15); bootstrap 95% CIs within $\pm 15\%$ (Table 5). The frequency column is the high-frequency energy ratio change (remover <0.9 / injector >1.1). Classes follow the pre-registered rule of Section 4.2.

Operator	α_{task}	α_{topo}	Freq. direction	Class
JPEG	0.71	0.63	remover	purifier
Gaussian noise	1.28	1.03	injector	masker
Coarse dropout	1.16	1.01	injector	masker
Block occlusion	0.64	0.54	injector	masker
Affine	1.48	0.95	injector	masker
Elastic	1.30	0.72	injector	masker
Blur	3.13	2.22	remover	over-remover
Downscale	6.15	4.92	injector	destructor
Grid dropout	33.22	10.96	injector	destructor

Table 2: Cross-domain consistency probes: task-decay rates (operating window). Blur flips from destructive to harmless; destructors are stable. The music column gives the audio analog's two-point log-rate on the normalized audio coordinate and its measured class; coordinates are domain-internal — compare classes, not absolute rates. Elastic is excluded on CIFAR-10 (knob miscalibration; see appendix).

Operator	Molecular α_{task}	CIFAR-10 α_{task}	Music analog
Blur	3.13	0.11	—
JPEG	0.71	2.42	0.16, purifier (mp3)
Gaussian noise	1.28	2.84	0.06, masker (white noise)
Coarse dropout	1.16	1.61	0.32, neutral (time dropout)
Downscale	6.15	1.96	0.02, purifier (lowpass)
Grid dropout	33.22	(lethal in both)	

(two-point log-rates; topology-destroying at $\alpha_{\text{topo}} \geq 2.0$, safe at ≤ 1.2; frequency remover below 0.9, injector above 1.1; full mapping in Appendix A).

Two cross-domain checks confirm this is a property of the domain, not of the operator: the frequency argument predicts the classification flips on CIFAR-10, whose foregrounds are low-frequency, and it does — blur's task-decay rate falls $28\times$ (from 3.13 to 0.11) while JPEG becomes *more* destructive ($0.71 \rightarrow 2.42$) as block artifacts attack coarse 32×32 structure (Table 2); and the same instrument transfers to a second oracle-bearing domain, music audio, where the measured classes again agree with the frequency argument (Appendix J).

4.3 What the measurement rules out

The signatures also reveal three qualitatively distinct decay morphologies — *threshold collapse* (downscale loses 50% of task performance at L=0.16, grid dropout at L=0.08), *gradual decay* (Gaussian noise, coarse dropout), and *near-harmless* behavior (affine, elastic, block occlusion) — and they rule out a family of theories. Destruction is threshold-like, not gradual: exponential decay families, the natural first choice for rate-based prediction, underfit the composite response severely ($R^2 \leq 0.28$; details in the appendix). Any predictor built on smooth rate extrapolation is therefore excluded before it is tested. What survives is deliberately simple: two-point log-rates over the operating window, composed multiplicatively, with landmarks located by a single survival threshold. Section 5 stress-tests exactly this estimator.

5 Stress-Testing a First-Order Transition-Landmark Estimator

We now test a deliberately first-order hypothesis — that a schedule's transition landmarks can be estimated from per-operator curves alone — letting the data arbitrate where it holds.

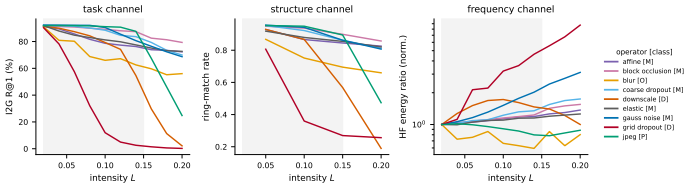

Figure 2: **Decay signatures.** Per-operator decay on the three channels (task, structure, frequency) over $L \in [0.02, 0.20]$, colored by measured class. The shaded region is the operating window. Threshold vs. gradual morphologies are visible; the frequency channel (log scale) separates removers (JPEG, blur) from injectors.

5.1 The estimator

Three rules, fixed before any held-out training: (i) composite topology survival is the product of member operators' measured ring-match curves (independence under sequential application); (ii) the critical threshold θ is the balanced reference schedule's composite survival at its observed optimum $L_{\text{opt}} = 0.152$, giving $\theta = 0.2961$, calibrated exactly once; (iii) for coupling variants, member curves are read at the effective knob intensity implied by the transform's saturation semantics. The predicted landmark of a schedule is the smallest L at which its composite survival crosses θ — the *cliff point*. On the reference schedule itself, the rule returns 0.1521 against the observed 0.152. All predictions, brackets, and verdict rules were pre-registered (Appendix A); a second prospective cohort, pre-registered with its protocol and predictions, is in the supplementary material — registered but not yet trained at submission, so the applicability rules of §5.3 remain hypothesis-generating until that cohort is run.

5.2 A boundary map, not a leaderboard

Ten held-out schedules constitute not a leaderboard but a boundary map (Table 3, Figure 3). The estimator is exact at bracket resolution where its assumptions hold: two point hits, including one exact to the grid argmax (N5, Δ=0.0000; Δ=0.002 under a fitted-vertex re-scoring, Appendix A), and one confirmed negative prediction (N7: no cliff predicted, none observed). Where it fails, it fails in three observed structural ways: *saturation* (V1, V2), where asymmetric coupling saturates one knob early and the transition region vanishes; *interaction error* (N3), where measured joint survival deviates from the product with sign flips no frequency-overlap rule predicts; and *compensability*, where anchor training learns to route around occluded patches (N3's remaining error term, and N8's undershoot). Two further schedules are false positives (N2, V3): a predicted cliff inside the probed range where none occurred — both are coupling variants that the saturation rule would have flagged out-of-scope before any training (§5.3). N2's early classification as a hit reflected an implementation defect in the knob mapping; under corrected semantics it is a false positive (Appendix A, note). One schedule (N6) is a non-case: its nominal bracket peak sits within plateau noise and its joint survival there (0.50) is far above the transition band — no landmark existed to predict. The pre-registered primary criterion — at least three of four within $|\Delta| \leq 0.010$ — is not met (2/4 under the plateau reading, 1/4 under a resolution-matched re-scoring; Appendix A); read under the pre-registration, this negative result is itself the boundary: it delimits where a first-order, independence-assuming estimator of degradation schedules can be trusted.

5.3 Why the failures fail

The failures decompose into three measurable boundary conditions. *First, saturation*: when a knob saturates inside the operating window — under asymmetric coupling one side saturates early, as in V1/V2 — the competition never develops and the response is a flat plateau. Saturation is sufficient for failure, not necessary for transition absence: at extreme coupling (N2, k=3) the purifier side saturates at $L \approx 0.05$, yet the response stays a flat, high plateau throughout the probed range. *Second, the product rule is a first-order approximation with measurable interaction error*: joint ring-match curves for operator pairs deviate from the product of single-operator curves in both directions, with sign flips even within one pair across intensity — the error is not even sign-consistent

Table 3: Pre-registered predictions vs. observations (deployment protocol, test L=0.150). Naive baseline: train at deployment intensity (0.150). Verdicts: *hit* = observed landmark within ± 0.010 of prediction (plateau reading allowed); *false positive* = predicted cliff inside probed range, no transition; *miss* = located transition beyond tolerance; *negative hit* = no cliff predicted or observed; *no true transition* = post-hoc non-case. Each schedule was registered before its own verification runs (Appendix A).

Schedule	Predicted	Observed	Verdict
N1 (composition)	0.1564	plateau $[0.156, 0.171]$	**hit** (plateau reading: pred. inside plateau; grid argmax Δ=0.0146)
N2 (3:1 coupling)	0.4595	flat 91.7–92.3 on $[0.445, 0.475]$	false positive
V1, V2 (coupling)	up, 0.31	flat plateaus	miss (saturation, §5.3)
N3 (noise+grid dropout)	0.1133	peak 0.1733 (91.98 ± 0.06)	**miss** (Δ=0.060; interaction + compensability)
N5 (noise+downscale)	0.1704	0.1704 (91.43 ± 0.30)	**hit** (Δ=0.0000 grid argmax; 0.002 fitted)
N6 (jpeg+downscale)	0.1656	plateau noise; joint survival 0.50 at nominal peak	no true transition (§5.3)
N7 (blur+noise)	no cliff	dip 0.050 (77.94 ± 2.36), then flat	**negative hit** (no cliff-and-recovery)
N8 (jpeg+grid dropout)	0.1164	peak 0.1464 (91.95 ± 0.23)	**miss** (Δ=0.030; undershoot)
V3 (1.5:1 coupling)	0.2298	flat 91.4–92.0 on $[0.215, 0.245]$	false positive

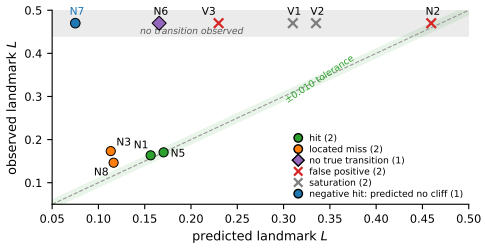

Figure 3: **Boundary map of the ten-schedule stress test.** Predicted vs. observed landmark per schedule; the band is the pre-registered ± 0.010 tolerance. False positives (N2, V3) and saturation cases (V1, V2) predicted a cliff where none occurred; N7's negative prediction was confirmed; N6's nominal peak is plateau noise — no true transition (joint survival 0.50, far above the band).

across schedules (Appendix I). The N3 miss decomposes into interaction error (0.027) and anchor training's occlusion-compensability (a further 0.033) (Appendix I). The same decomposition applied symmetrically to the hit: N5's joint curve crosses θ at 0.156 — 0.015 earlier than the product-based prediction — while the observed landmark sits exactly at the prediction, so N5's exactness rests on a near-cancellation of interaction error (-0.015) and compensability ($+0.015$), not on the absence of either term (Appendix I). *Third, compensability*: occlusion-type destruction is partly routable, so structure survives beyond what the reference-model curves record; N8's joint curve is super-multiplicative at high L ($+57\%$ at 0.18, $+121\%$ at 0.20), and its observed landmark sits beyond both the product and joint crossings, the same interaction-plus-compensability decomposition as N3 (Appendix I). N6, by contrast, had no transition to miss: its response is a plateau within noise and joint survival at the nominal peak (0.50) never approaches θ.

Two applicability remarks. The saturation rule — derivable from the knob mapping alone — flags all four coupling schedules (V1–V3, N2) as out-of-scope without any training. Its sufficiency was derived from the observed flat responses it now explains; the same post-hoc status applies to all three boundary conditions of this section — hypothesis-generating here, genuinely prior only for the prospective cohort. The threshold is approximately schedule-invariant: measured *joint* survival at the observed landmarks clusters at 0.22–0.29 across schedules (Appendix I) — the weak link is the product-independence assumption, not the threshold. L_{opt}'s calibration uncertainty (± 0.005) drifts predictions by ± 0.004 (N1) to ± 0.035 (N3), growing with extrapolation distance.

5.4 WHAT SURVIVES: DEPLOYMENT-CONDITIONAL ANCHOR SELECTION

The strongest naive baseline is *train at the deployment intensity*. It is point-optimal: on the reference schedule, the naive anchor at $L{=}0.150$ reaches 93.15 ± 0.57 at its design test — far above every alternative. But the deployment answer changes with range (Appendix Table 8). The zero-cost MixTrain baseline — uniform random intensity per sample, the RandAugment-family default — dominates the mid range and beats the high-intensity anchor at test 0.200 by 52.01 to 45.00. Only in the extreme tail does the high-intensity anchor lead: $+13.8$ at test 0.250 and $4.8\times$ at 0.300, where MixTrain itself collapses. Two provenance disclosures: the deep anchor at 0.450 was selected post-hoc, so the table shows deep anchoring *can* pay off in the tail — not that the predictor chose it; and all rows are reference-schedule runs under the deployment protocol (the deep and MixTrain rows partly predate the coupling-semantics correction; verified against training logs).

The decision rule, stated explicitly. A practitioner with a *known* deployment point should train at that intensity — the naive rule is point-optimal. Facing an *unknown or moderate* range, use mixture training. Only in the *extreme tail*, or when designing *many* schedules in one coordinate system, does the measurement-plus-predictor route pay off: measure the operators, locate the cliff, and step back one trainability margin.

The trainability margin qualifies this further: a post-hoc 5-seed fine scan of the plateau edge (test $L{=}0.150$) shows run-level instability at the boundary (std $\leq$ 0.8 across the plateau but 4.2 at $L{=}0.160$) — deploy toward the plateau interior, not the edge (here: 0.152–0.156, not 0.160–0.162; cost accounting in Appendix I: breakeven at $\gtrsim$9 schedules).

Generalization to held-out scaffolds. Under a Bemis–Murcko scaffold split (Bemis & Murcko, 1996) (strictly disjoint 193k/2k/5k), the mild-anchor penalty and the deployment-conditional optimum replicate with undiminished magnitude (appendix Table 10): at test $L{=}0.152$, L_{opt} exceeds the texture-apex anchor by $+6.3$ points OOD against $+7.2$ ID; at $L{=}0.200$, $+20.5$ against $+23.6$ ID. Both also replicate across backbones and graph encoders (appendix).

6 DISCUSSION AND LIMITATIONS

The structure-oracle requirement. Why did the phenomenon surface here, not on standard image benchmarks? The answer is a three-part *structure oracle*, not "having ground truth": **discreteness** — edits are atomic; **legality grammar** — an edit yields another valid molecule or is invalid, with no in-between; and **canonical representation** — a graph completely describes structure, so ring-match is a calibrated ruler. Natural images lack all three. This defines an oracle-quality axis — code, proofs, text, audio — with the music-audio probe (Appendix J) mid-axis. The asymmetric design — degraded images against a clean graph anchor — is the second load-bearing choice: the clean target is what makes structure survival measurable at all.

Limitations (stated as precisely as the results). (i) L is an operational coordinate: predictions hold within one coordinate system and knob mapping; the product rule assumes independent structure effects. (ii) The estimator locates *landmarks*, not the deployment optimum — the decision rule of §5.4 applies. (iii) It fails for compensable occlusion-type destruction — a ceiling, not a gap: compensability is a property of what anchor training *learns* to route around — invisible in principle to pre-training measurement — and five of nine operators are occlusion- or injection-type. Absolute survival values shift with the reference model, though classes and fixed-θ predictions are stable (Appendix C). (iv) Validation covers one molecular retrieval task plus measurement-layer probes on CIFAR-10 and music audio. (v) Ring-match verifies ring topology but is blind to atom types, bond orders, and stereochemistry, so the structure channel underestimates chemical fidelity.

What this opens. The boundary map is a beginning, not a verdict — its three failure classes are the research agenda: interaction error calls for measured joint curves, saturation is a provable scope rule, and compensability defines what a second-order estimator must model. The pre-registered second cohort tests these rules (Appendix B): all five no-cliff predictions hold, the gate is correct on $3/4$ and falsified once, as registered (K3), and point predictions remain first-order-fragile ($1/3$, the one hit at the tolerance edge).

AI USE STATEMENT

AI-assisted tools were used for literature retrieval and discovery, research ideation, drafting assistance, and code editing during the preparation of this work. All experimental results were produced by the authors' own pipelines; every reported number was verified against raw experiment logs by the authors. The authors take full responsibility for the content.

REPRODUCIBILITY STATEMENT

All results are reproducible from the released materials: training and evaluation code (including the single-operator scan, composite-schedule training, topology-preservation, and frequency-signature scripts), the raw spectra CSVs underlying every figure and table, the pre-registered prediction files with in-file timestamps, per-seed logs for all reported runs, and the exact train/val/test splits (including the scaffold split). The dataset is derived from the public PubChem repository; processing scripts are included.

REFERENCES

Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *NeurIPS*, 2019.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

Simone Bombari and Marco Mondelli. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In *ICML*, 2025.

Abdelali Bouyahia, Frédéric LeBlanc, and Mario Marchand. The geometry of invariant learning: An information-theoretic analysis of data augmentation and generalization. *arXiv:2602.14423*, 2026.

Haochen Chen, Syed Fahad Sultan, Yingtao Tian, Muhao Chen, and Steven Skiena. Fastrp: Fast and accurate network embeddings via random projections. In *CIKM*, 2019.

Shuxiao Chen, Edgar Dobriban, and Jane Lee. A group-theoretic framework for data augmentation. In *NeurIPS*, 2020.

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *NeurIPS Workshop*, 2020.

Jingyi Cui, Hongwei Wen, and Yisen Wang. An augmentation-aware theory for self-supervised contrastive learning. *arXiv:2505.22196*, 2025.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *ICML*, 2019.

Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent. In *NeurIPS Workshop on Mathematics of Modern Machine Learning*, 2023.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. In *ISMIR*, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Raphael Gontijo-Lopes, Sylvia J. Smullin, Ekin D. Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv:2006.06289*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

Daniel Ho, Eric Liang, Xi Chen, Ion Stoica, and Pieter Abbeel. Population based augmentation: Efficient learning of augmentation policy schedules. In *ICML*, 2019.

Yi Huang et al. Is the information bottleneck robust enough? towards label-noise resistant information bottleneck learning. In *AAAI*, 2026.

Yu Huang et al. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. In *COLM*, 2024.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. In *ICML*, 2024.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *NeurIPS*, 2022.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *ICML*, 2023.

Polina Kirichenko, Mark Ibrahim, Randall Balestriero, Diane Bouchacourt, Ramakrishna Vedantam, Hamed Firooz, and Andrew Gordon Wilson. Understanding the detrimental class-level effects of data augmentation. In *NeurIPS*, 2023.

James B. Lin, Divyansh Kaushik, Ethan Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *JMLR*, 25(91): 1–85, 2024.

Ce Liu, William T. Freeman, Richard Szeliski, and Sing Bing Kang. Noise estimation from a single image. In *CVPR*, pp. 901–908, 2006.

Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Single-image noise level estimation for blind denoising. *IEEE TIP*, 22(12):5226–5237, 2013.

Z. Liu et al. Molecular motif learning as a pretraining objective for molecular property prediction. *Nature Communications*, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *ICLR*, 2023.

Sachin Mehta, Saeid Naderiparizi, Fartash Faghri, Maxwell Horton, Lailin Chen, Ali Farhadi, Oncel Tuzel, and Mohammad Rastegari. Rangeaugment: Efficient online augmentation with range learning. In *ICLR*, 2023.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *NeurIPS*, 2021.

Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *ICCV*, 2021.

Chitwan Mummadi, Akshay Subramaniam, et al. Does enhanced shape bias improve neural network robustness to common corruptions? In *ICLR*, 2021.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *ICLR*, 2020.

Dang Nguyen et al. Changing the training data distribution to reduce simplicity bias improves in-distribution generalization. *arXiv:2404.17768*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, 2021.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*, 2022.

Bowen Qiu, Zhangir Kuang, and Surbhi Goel. Complexity matters: Dynamics of feature learning in the presence of spurious correlations. In *ICML*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Evgenia Rusak, Lukas Schneider, Amir Zamir, et al. A simple way to make neural networks robust against diverse image corruptions. In *ICLR*, 2022.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020.

Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation. In *ICML*, 2022.

Samarth Sinha and Animesh Garg. Curriculum by smoothing. In *NeurIPS*, 2020.

Teppei Suzuki. Teachaugment: Data augmentation optimization using teacher knowledge. In *CVPR*, 2022.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020.

Khanh Trinh, Markus Heinonen, Luigi Acerbi, and Samuel Kaski. Damp: Improving robustness to corruptions with multiplicative weight perturbations. In *NeurIPS*, 2024.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *NeurIPS*, 2021.

Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, et al. Augmax: Adversarial composition of random augmentations for robust training. In *NeurIPS*, 2021.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279–287, 2022.

Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *ICIP*, pp. 477–480, 2002.

Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *ICML*, 2021.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

Suorong Yang, Furao Shen, and Jian Zhao. Entaugment: Entropy-driven adaptive data augmentation framework for image classification. In *ECCV*, 2024.

Suorong Yang, Peijia Li, Xin Xiong, Furao Shen, and Jian Zhao. Adaaugment: A tuning-free and adaptive approach to enhance data augmentation. *IEEE TIP*, 34:4843–4855, 2025.

Suorong Yang, Hongchao Yang, Suhan Guo, Furao Shen, and Jian Zhao. Ipf-rda: An information-preserving framework for robust data augmentation. *IEEE TPAMI*, 48(2):1339–1354, 2026.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning with augmentations. In *NeurIPS*, 2020.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021.

A Pre-registered classification rule and prediction protocol

Classification rule (verbatim from the pre-registration). Decay rates are two-point log-rates $\alpha = \ln(y_{L_1}/y_{L_2})/(L_2 - L_1)$ over the operating window (task channel: $L \in [0.02, 0.15]$; structure channel: $L \in [0.05, 0.15]$), which is functional-form free. The frequency direction is the ratio of the high-frequency energy ratio at L=0.15 vs. L=0.02: remover (< 0.9), injector (> 1.1), else neutral. The structure channel is dominant: $\alpha_{\text{topo}} \geq 2.0 \rightarrow$ topology-destroying; $\alpha_{\text{topo}} \leq 1.2 \rightarrow$ topology-safe; between: ambiguous. Final classes: (remover, safe) $\rightarrow$ purifier; (injector, safe) $\rightarrow$ masker; (neutral, safe) $\rightarrow$ neutral; (any, destroying) $\rightarrow$ destructor, with (remover, destroying) labeled over-remover. Exact rates are reported in Table 1; the decisive contrasts are threshold-insensitive.

Topology-cliff predictor (verbatim rules). (i) Composite topology survival = product of member operators' measured ring-match curves. (ii) θ = composite survival of the balanced reference schedule at its observed optimum $L_{\text{opt}} = 0.152$ ($\theta = 0.2961$; calibrated exactly once). (iii) Coupling: strengthened side capped at knob ratio 1, weakened side uncapped (matching the transform implementation); balanced/subset schedules uncapped. (iv) Predicted L_{opt} = smallest L with composite survival $\leq \theta$ (scan $L \in [0.02, 0.60]$). (v) Naive baseline: $L_{\text{opt}} = 0.152$ for all schedules. (vi) Tolerance: $|\Delta| \leq 0.010$. (vii) Out-of-range rule: predictions beyond the verifiable range are marked extrapolated; a monotone observed increase through the trained range counts as a directional hit.

Pre-registered scorecard (verbatim). Primary: $\geq 3/4$ schedules with $|L_{\text{opt}}^{\text{pred}} - L_{\text{opt}}^{\text{obs}}| \leq 0.010$. Secondary: coupling shift direction correct $\geq 3/4$; model beats naive on $\geq 3/4$. All outcomes reported, including misses. *Outcome: primary 2/4 (not met). Extended to ten schedules (six pre-registered, four added during revision, each registered before its own verification run): point hits 2/9 (N1 plateau-limited, N5 exact); false positives 2 (N2, V3 — predicted cliffs lie inside the probed range, no transition observed); negative prediction 1/1 (N7); directional (V1, V2) 0/2 (saturation); located misses N3, N8 (interaction + compensability); N6 reclassified post hoc as no true transition (plateau noise; joint survival* 0.50 *at the nominal peak, far above the transition band).*

Resolution-matched re-scoring. The verification brackets of the point-prediction schedules are spaced at 0.015 — coarser than the pre-registered tolerance (± 0.010) — and were centered on the predictions by protocol design, so a grid-argmax scorecard is near-binary and N5's "Δ=0.0000" certifies only that the argmax fell on the center point. We therefore re-score the primary cohort under a resolution-matched criterion, $|\Delta| \leq$ half the grid spacing, with the observed optimum located by a quadratic vertex fit through the peak neighborhood. **N5 survives**: fitted argmax 0.1721 ($\Delta = 0.0017$; per-seed fits 0.170–0.176 where well-conditioned) — though the peak's contrast over the flanks is $+0.84$/$+0.53$ R@1 (paired $t \approx 1.6$, n.s. at n=3), so the hit is real at bracket resolution but the peak is shallow. **N1 is verdict-fragile**: its response is a plateau (paired 0.156 vs. 0.171: $+0.41$,

s.e. 0.48, n.s.), so the prediction sitting inside the plateau is a hit under the plateau reading, while the grid argmax (0.171, $\Delta = 0.0146$) misses under both tolerances. N2 (false positive) and N3 ($\Delta \geq 0.060$; response still rising at the bracket edge) are unchanged. Re-scored primary: $2/4$ under the plateau reading, $1/4$ under strict argmax — **not met under either**: the negative headline is not a bracket-resolution artifact. Bracket spacing for the prospective cohort is tightened to ≤ 0.005.

Knob-mapping note. An early version of the predictor capped all member curves at ratio 1; this was corrected to the transform's actual semantics (strengthened-side-only cap) *before* pre-registration was finalized. The sanity check (balanced $\rightarrow$ 0.1521) passed under the corrected mapping.

Implementation-defect note (N2). Separately, the *training-side* coupling knob was inert in the first verification vintage (a mapping bug): N2's early runs were de facto balanced-schedule trainings, and its initial "hit" verdict reflected that defect. All N2 runs were retrained under the corrected semantics (flat plateau across $[0.445, 0.475]$; Table 6), and the verdict was reclassified from hit to false positive. The audit trail (logs, checkpoints, and the revision audit paragraph of Appendix F) is preserved in the supplementary material.

Registration timeline. What existed when, stated plainly. *Fixed at or before each batch's registration*: rules (i)–(vii) including the knob-mapping cap correction and the θ calibration (`PREDICTIONS.md`, 2026-08-05, before any held-out training; schedules N1, N2, V1, V2); batch two (`PREDICTIONS_N3N5.md`, 2026-08-13; N3, N5, before their training); batch three (`PREDICTIONS_BATCH2.md`, 2026-08-20; N6–N8, V3, each before its own training — note the *choice* of these four was informed by batch-one/two outcomes: they were selected to be discriminative, and registration covers each schedule's prediction, not that selection). *Derived after observing failures* (post-hoc, and labeled as such wherever used): the saturation scope rule of §5.3, N6's no-true-transition criterion (joint survival at the nominal peak far above θ), the trainability margin, and the 0.450 deep anchor.

Secondary criteria outcomes. The scorecard pre-registers two secondary criteria; both are reported here with the same fidelity as the primary. *Coupling shift direction* $\geq 3/4$: **not met**. All four coupling schedules (V1, V2, N2, V3) returned flat plateaus — no transition occurred, so no shift direction exists to verify ($0/2$ at batch one; $0/4$ including the post-correction N2 and batch-three V3). *Beats naive* $\geq 3/4$: **not met**. At the batch-one verification vintage this read $2/2$ (N1, N2), but N2's runs were later invalidated by the knob defect, and N1's prediction (0.1564) nearly coincides with the naive anchor (0.152) — non-discriminative by design, as flagged at registration. Among held-out-operator schedules, the predicted anchor beats the naive-adjacent bracket point on N5 (91.43 at 0.1704 vs. 90.59 at 0.1554) and loses on N3 (88.27 at 0.1133 vs. 91.37 at 0.1583).

B Prospective second cohort: the applicability rules tested under pre-registration

Design and registration. The second cohort comprises twelve schedules — eight class-stratified operator pairs (C1–C8) and four coupling variants of the composite (K1–K4) — registered on 2026-09-03 (`PREDICTIONS_BATCH3.md`), amended once before training (A1, 2026-09-11: five-point verification brackets at 0.005 spacing). Verification ran 131 training runs (the registered 126 plus protocol-triggered edge extensions and 5-seed top-ups; zero failed runs) under the same deployment protocol (test L=0.150). Verdicts were computed by a fixed script against the frozen scorecard (hit: $|\Delta| \leq 0.010$ with an interior, resolvable peak; false positive / false negative / censored as registered). All per-point means, per-seed values, and the scoring script ship in the supplementary material (`stage26/VERDICTS.md`).

What the cohort adds. Three findings. (i) *The negative side and the gates are prospectively reliable*: all five no-cliff predictions hold and the saturation gate correctly idles K1, K2, K4 — applicability classification $9/10$ decidable schedules ($9/12$ counting the two censored); C7 shows a $\pm$2-point bump at 0.150, below the operationalized materiality floor (5 R@1 points; `VERDICTS.md`). (ii) *The protocol catches its own false alarms*: C2's three-seed bracket suggested a clean hit (peak

Table 4: Second-cohort verdicts (deployment protocol, test L=0.150; 3 seeds per point, 5 where marked). Per-point curves and per-seed values: supplementary `VERDICTS.md`.

Schedule	Registered	Observed	Verdict
C2 (destructor×destructor)	0.1065	peak 0.1115 (63.9 ± 14.8, n=5), unresolvable	censored
C3 (masker×destructor)	0.1150	interior peak 0.1250 (91.00 ± 0.08)	**hit** ($\|\Delta\|$=0.010)
C8 (masker×destructor)	0.1120	plateau 89.0–90.1 (six points incl. 0.1270 ext.)	censored
C1, C4–C7	no cliff	flat/monotone, 89–92	negative hit ×5
K1 (k=1.5)	out-of-scope (saturation)	flat 91.4–92.0	gate correct
K2 (k=2.5)	out-of-scope (saturation)	flat 91.6–92.2	gate correct
K3 (k=0.5)	out-of-scope (saturation)	transition at 0.150 (46.3→61.5→29.8)	**gate falsified**
K4 (k=4.0)	out-of-scope (saturation)	flat 92.0–92.3	gate correct

0.1115, Δ=0.005), but the registered 5-seed top-up flipped it — 63.9 ± 14.8 across five seeds, no turning point resolvable above seed noise. The landmark region is correct, but run-level response at a destructor×destructor cliff is chaotic — an extreme instance of the boundary instability of §5.4. (iii) *The gate is falsifiable, and was falsified*: K3 (k=0.5) shows a transition at 0.150 (46.3 → 61.5 → 29.8, all three seeds) where the saturation rule registered "flat, out of scope", localizing the rule's blind spot to the sub-unitary coupling regime; one seed additionally collapsed entirely at L=0.200 (R@1 = 0.02), reported per the registered failure discipline.

Point predictions and the three estimators. C3 is a borderline hit: the registered edge-expansion rule added 0.1300, whose roll-off (90.21) made 0.1250 an interior peak — $|0.1250 - 0.1150| = 0.010$, exactly at the tolerance boundary, with a small effect (0.4–0.8 points) but seed-tight (±0.08). The applicable-subset tally is $1/3$ (Wilson 95% CI $[0.061, 0.792]$), point MAE 0.0100 (n=1), false positives 0, FPR $1/9$. Replacing the product assumption with measured joint curves (third estimator) improves C3 (0.1280 predicted vs. 0.1250 observed) but fails identically on C2/C8 and misses K3 — the bottleneck is not the product assumption.

C Measurement protocols

Reference model. All Stage-0 spectra use the clean-trained baseline checkpoint (degradation-naive), so signatures characterize operators, not trained robustness.

Reference-model stability. Because ring-match is computed through the retrieval model, signatures are task-level behavioral measurements, not operator-intrinsic constants. As a robustness check we re-measured *all five predictor-relevant operators* (blur, JPEG, downscale, Gaussian noise, grid dropout; full curves at 0.01 spacing) under a second, independently trained reference checkpoint. Classes are unchanged for every operator — e.g. downscale $\alpha_{\text{topo}} = 4.92$ vs. 4.86, blur 2.22 vs. 2.90 — but absolute curves shift, most for blur (-20 to -26% ring-match across the window). Propagating the shift through the predictor: θ re-calibrated on the second reference moves from 0.2961 to 0.2078 (-30%), and with θ transferred *without* recalibration the predicted cliffs of all six evaluated schedules move by at most 0.012 (-0.012 to $+0.002$), changing no verdict; with per-reference recalibration the shifts are larger (up to $+0.030$) and move the two located misses (N3, N8) *toward* their observed optima. We therefore read signatures as stable at the level of classes, orderings, and fixed-θ landmark transfer — not as exact constants.

Static-cache note. Evaluations bypass the pre-generated static benchmark cache (`--static_bench_dir` pointed at a non-existent path) because the cache covers only the original test split's molecule IDs; all reported numbers use dynamically drawn degradations with fixed draws across checkpoints at each (operator, L).

CIFAR-10 protocol. Task channel only: frozen ImageNet-ResNet18 features + logistic probe trained on clean CIFAR-10 train split; 500 test images; operators reimplemented with identical knob formulas (image-library version differences mapped explicitly). Elastic is excluded: the local transform library's elastic implementation uses a different parameterization that destroys images at every L (chance-level accuracy throughout); this is an environment artifact, not a scientific finding. The structure channel (low-frequency band cosine) saturated on CIFAR-10 and is not reported. Cross-domain comparison is rate-level, not accuracy-level.

D TRAINING CONFIGURATION AND REPRODUCIBILITY

Architecture. CLIP-style dual tower: ConvNeXt-Base vision encoder and a 5-layer GIN over FastRP graph embeddings, projected to a shared 512-d space. Contrastive logits use a learnable logit scale (temperature $= 1/\text{logit-scale}$, clamped at 100). Loss: InfoNCE with hard-negative mining (margin 1.1, push weight 1.0).

Optimization. AdamW ($\text{lr} = 10^{-3}$, weight decay 10^{-4}), CosineAnnealingLR ($T_{\max} = \text{epochs}$), batch size 256, 10 epochs, mixed precision (AMP). Model selection: best validation R@1 at the fixed design test L=0.150 (validation split of 2,000 held out from the training pool). All timings on a single RTX 5090.

Data. 200,000 PubChem image–graph pairs, split 193k/2k/5k (train/val/test); retrieval gallery at evaluation is the 5,000-member test split. Scaffold split: 89,827 Bemis–Murcko scaffolds, strictly disjoint partitions.

Degradation coordinate and knob mapping. Intensity ratio $r = L/0.150$ (balanced: both sides at ratio r; coupling k: strengthened side $\min(kr, 1)$, weakened side r/k). Composite operators, all applied with p=1: blur = GaussianBlur, limit $(3,\ \max(3, \lfloor 3 + 6r \rfloor))$ rounded up to odd; JPEG = ImageCompression, quality $(\max(1, \lfloor 100 - 90r \rfloor),\ \max(10, \lfloor 100 - 70r \rfloor))$; downscale = Downscale, scale $(\max(0.05, 1 - 0.7r),\ \max(0.1, 1 - 0.5r))$; coarse dropout = CoarseDropout, holes 1–$\max(1, \lfloor 5r \rfloor)$, hole size 5–$\max(5, \lfloor 35r \rfloor)$ px, fill 0. Subset composites (N1–N8) reuse these per-operator formulas on their member operators.

Run counts and seeds. Valley scan: 5 seeds (42, 1111, 1234, 2026, 2027). Held-out schedules and baselines: 2–3 seeds (42, 1234, 2026) as marked per table. Evaluation draws are fixed per (operator, L) across all checkpoints.

Evaluation protocols (deployment vs. self-matched). Our primary criterion is the deployment protocol (best val R@1 at fixed extreme test condition, test L=0.150), consistent with the operational definition of the optimal anchor. Self-matched evaluation (test L = training L) is used only to trace the U-curve itself (Figure 1 and the boundary plateaus); it conflates model quality with test difficulty across schedules and is reported only as a diagnostic. The training-side U-curve of Figure 1 (anchors scanned along training intensity, self-matched test) is a different object from the test-side dip observed when a fixed extreme anchor is evaluated across test intensities; the two are related through the anchor's specialization range and should not be conflated.

E FULL MEASURED RATES

Operating-window and full-grid task rates, structure rates, and frequency signatures per operator are reported in Table 1 (main text) and in machine-readable form in the supplementary CSVs (`spectra_task.csv`, `spectra_topo.csv`, `spectra_freq.csv`, 9 operators × 10 intensities; fine structure spectra at 0.01 spacing for the four composite operators). Full-grid rates (used as morphology diagnostics only): blur 2.64, JPEG 7.10, downscale 19.98, grid dropout 31.33, Gaussian noise 1.57, coarse dropout 1.38, affine 1.31, elastic 1.32, block occlusion 0.85.

Bootstrap uncertainty and endpoint correction. All rates are computed from per-molecule records (5,000 test molecules) at exact window endpoints; Table 5 reports paired-bootstrap 95% CIs (1,000 resamples over molecules). Two notes. First, an earlier version of Table 1 interpolated the task channel at L=0.15 (the task grid lacked that point); all rates now use exact measured endpoints, shifting three task-channel values (JPEG 1.31 → 0.71, coarse dropout 0.94 → 1.16, grid dropout 29.27 → 33.22); no class, and no downstream claim, changes. Second, the CIs capture molecule-sampling variability; degradation-draw variability is small and not included.

Table 5: Bootstrap 95% CIs for the measured decay rates (1,000 paired resamples over the 5,000 test molecules). $\alpha_{\text{task}}^{\text{full}}$ is the full-grid task rate ($L \in [0.02, 0.20]$), a morphology diagnostic. Classification-relevant margins: every operator's class is stable across its CI (e.g. blur $\alpha_{\text{topo}} \in [2.05, 2.42]$ vs. the destructor threshold 2.0).

Operator	α_{task} [95% CI]	α_{topo} [95% CI]	$\alpha_{\text{task}}^{\text{full}}$ [95% CI]
JPEG	0.71 [0.64, 0.79]	0.63 [0.56, 0.71]	7.10 [6.83, 7.35]
Gaussian noise	1.28 [1.18, 1.37]	1.03 [0.93, 1.13]	1.57 [1.47, 1.66]
Coarse dropout	1.16 [1.06, 1.26]	1.01 [0.91, 1.11]	1.38 [1.28, 1.46]
Block occlusion	0.64 [0.57, 0.72]	0.54 [0.46, 0.63]	0.85 [0.78, 0.92]
Affine	1.48 [1.38, 1.59]	0.95 [0.84, 1.07]	1.31 [1.22, 1.39]
Elastic	1.30 [1.20, 1.40]	0.72 [0.61, 0.83]	1.32 [1.23, 1.41]
Blur	3.13 [2.96, 3.29]	2.22 [2.05, 2.42]	2.64 [2.52, 2.78]
Downscale	6.15 [5.91, 6.41]	4.92 [4.70, 5.16]	19.98 [19.04, 21.00]
Grid dropout	33.22 [31.39, 35.52]	10.96 [10.51, 11.45]	31.33 [28.87, 35.15]

F PER-RUN VERIFICATION TABLES

Table 6 reports every verification run for the eight bracketed schedules (deployment protocol, test L=0.150), including prediction source files, bracket definitions, and per-seed values.

Revision audit trail. Several reported numbers changed across revision; each change is one of three kinds — protocol unification, an implementation correction, or a data correction — and verdicts are unchanged unless stated: (i) *N1, N5 bracket values* moved by ∼1–5 points because pre-revision values used the validation protocol (2k gallery, best-val selection) while the revision reports the deployment test protocol (5k gallery) throughout; (ii) *N2* was retrained under the corrected coupling-knob semantics (the pre-revision implementation misread the mapping), and its verdict changed from hit to false positive; (iii) *N3* had two race-contaminated checkpoints (L=0.1133, 0.1283, seed 42) retrained and re-evaluated — the observed peak is unchanged at 0.1733; (iv) *Figure 1b*: three contaminated clean-baseline checkpoints were retrained, moving the baseline annotation from 58.3 to 13.7; (v) *Table 8*: the naive row was re-evaluated on three fine-scan checkpoints and the MixTrain row extended to three seeds, all under the deployment protocol; the deep-anchor row is unchanged.

N1 (2P1D; prediction 0.1564). Observed optimum: plateau $[0.156, 0.171]$; prediction inside the plateau ($\Delta_{\text{nearest}} = 0.0004$).

N2 (3:1 coupling; prediction 0.4595). Under the corrected coupling semantics, the probed range $[0.445, 0.475]$ shows a flat plateau (91.7–92.3, n=3 seeds each) with no transition. The predicted cliff lies inside the probed range and is not observed: a false positive. The k=3 coupling keeps both sides near the balanced optimum ($L_{\text{eff}} \approx 0.15$) throughout.

V1/V2. Flat plateaus (self-matched R@1): V1 peaks 94.72 at L=0.03 and ends 94.08 at L=0.20; V2 peaks 94.34 at L=0.06 and ends 93.48 at L=0.20. No valley, no peak across the trained range.

G FAILED PARAMETRIC MODEL (NEGATIVE RESULT, DETAILS)

The dual-exponential model $R(L) = U_s e^{-\alpha_p L} + U_t(e^{\alpha_c L} - e^{\alpha_d L})$ was fit on the balanced composite valley curve. Free 5-parameter fit: $R^2 = 0.278$. Anchored fit (α_p, α_d fixed to measured values; only U_s, U_t, α_c free): $R^2 = 0.114$, with U_t on its constraint boundary (standard error 194.8) and $\alpha_c \approx \alpha_d$ (near-degenerate). The model family cannot express threshold-type morphology; we do not use it anywhere in the paper.

Table 6: Per-seed verification runs for all bracketed schedules (deployment protocol, test L=0.150). Predictions were registered before each schedule's verification training: batch one (N1, N2, V1, V2) in `PREDICTIONS.md`, batch two (N3, N5) in `PREDICTIONS_N3N5.md`, and the revision-added four (N6–N8, V3) in `PREDICTIONS_BATCH2.md`; all timestamped files are in the released repository. Schedule labels follow the design-time order and are not contiguous (no N4 was registered).

Schedule	L	seed 42	seed 1234	seed 2026	mean ± std
N1	0.141	90.34	–	–	90.34 ± –
	0.156	90.42	91.70	91.08	91.07 ± 0.64
	0.171	91.60	91.22	91.60	91.47 ± 0.22
N2	0.445	92.54	92.34	90.60	91.83 ± 1.07
	0.45	92.14	91.58	92.58	92.10 ± 0.50
	0.455	92.46	92.30	91.72	92.16 ± 0.39
	0.46	91.82	91.82	91.40	91.68 ± 0.24
	0.465	92.56	91.94	92.46	92.32 ± 0.33
	0.47	92.34	90.90	92.36	91.87 ± 0.84
	0.475	92.40	92.42	91.96	92.26 ± 0.26
N3	0.0983	75.18	82.78	–	78.98 ± 5.37
	0.1133	87.76	88.78	–	88.27 ± 0.72
	0.1283	90.96	91.34	–	91.15 ± 0.27
	0.1433	91.84	91.66	–	91.75 ± 0.13
	0.1583	90.90	91.84	–	91.37 ± 0.66
	0.1733	91.94	92.02	–	91.98 ± 0.06
N5	0.1554	90.74	91.54	89.48	90.59 ± 1.04
	0.1704	91.74	91.40	91.14	91.43 ± 0.30
	0.1854	90.56	91.22	90.92	90.90 ± 0.33
N6	0.1206	90.46	88.72	90.54	89.91 ± 1.03
	0.1356	91.46	90.60	91.16	91.07 ± 0.44
	0.1506	91.48	91.78	92.08	91.78 ± 0.30
	0.1656	90.78	91.04	91.68	91.17 ± 0.46
	0.1806	89.86	90.72	91.40	90.66 ± 0.77
N7	0.05	79.24	79.36	75.22	77.94 ± 2.36
	0.1	92.46	91.62	92.58	92.22 ± 0.52
	0.152	92.32	92.32	92.44	92.36 ± 0.07
	0.2	91.88	92.16	92.56	92.20 ± 0.34
	0.3	92.36	91.98	91.50	91.95 ± 0.43
	0.45	90.52	92.18	90.58	91.09 ± 0.94
N8	0.1014	85.98	86.04	84.18	85.40 ± 1.06
	0.1164	88.30	88.24	89.40	88.65 ± 0.65
	0.1314	91.08	91.10	91.24	91.14 ± 0.09
	0.1464	92.20	91.74	91.92	91.95 ± 0.23
	0.1614	91.82	92.14	91.44	91.80 ± 0.35
V3	0.2148	91.94	91.88	92.22	92.01 ± 0.18
	0.2298	90.60	92.42	91.28	91.43 ± 0.92
	0.2448	91.68	91.88	91.70	91.75 ± 0.11

H DISCRETE-SCHEDULE COMPANION RESULTS

This section collects results on the legacy *discrete 5-level* schedule (clean, mild, intermediate, strong, extreme) used in the predecessor study; these levels are distinct from the continuous intensity coordinate L used throughout the paper.

Backbone generality. The non-monotonic transition and the anchor advantage replicate across ResNet-50, ConvNeXt-Base, and Swin-Base on the discrete 5-level schedule (3 seeds each; ConvNeXt extreme-anchor 88.60 ± 0.45 vs. best baseline 13.41 at the extreme level; full per-backbone tables in the supplementary CSVs).

Matched corruption-trained baselines (discrete schedule). MixTrain (uniform mixture of all levels, 2 seeds) recovers robustness (clean/strong/extreme levels $= 92.57/88.06/87.28$) but shows a *monotone* response — no U-shape in either seed; a frequency-matched control collapses (16.42 at the extreme level). At the extreme level the fixed anchor (88.60 ± 0.45, 3 seeds) exceeds both MixTrain seeds individually; at the intermediate level MixTrain wins (88.06 vs. 83.94) — a specialization trade-off we report openly.

Single-family scans (discrete schedule). The dip-and-recover survives within the downscale family alone (3 seeds: dip 86.27 ± 0.73 at s=0.45 $\rightarrow$ recovery 90.79 ± 0.13 at s=0.25, $+4.53 \pm 0.80$, paired $p = 0.010$; vanilla collapses monotonically to 2.26), is absent on blur, and is weak on JPEG — the transition appears precisely where structural destruction coexists with residual texture.

Table-4 anchor upgrade. The dip of the non-monotonic curve (strong-level training) was upgraded to 3 seeds: 82.59 ± 0.18 at extreme test vs. rebound 88.60 ± 0.45 — a 6.01-point dip, $p < 0.001$ (Welch $t \approx 21.5$).

MNIST texture-free control. Included in the supplementary CSVs; it replicates the published version (topology-retrieval negative control on MNIST).

Graph-encoder ablations. FastRP vs. Morgan fingerprints vs. random vs. symmetric-degradation controls, and auxiliary-module ablations (hard-negative mining, MMD, MoCo), are tabulated in the supplementary CSVs.

Baseline matrix. The discrete-schedule baseline matrix (VIB, DIB, MAE-Mask, and vanilla dual-tower, each against the extreme anchor across the five levels) is tabulated in the supplementary CSVs. Two custom baselines from the original study (PhysAug, InputMask) are omitted: their implementations could not be recovered for verification, and we report only baselines whose training code we can account for. An adapted AugMix baseline on the continuous framework is reported in Table 7.

Table 7: Adapted AugMix on the continuous framework (3 seeds) vs. the anchors, deployment protocol (fixed test L; anchor rows are 3-seed means $\pm$ std). AugMix = 3 chains, Dirichlet mixing; the original JSD consistency is unstable on retrieval logits and is replaced by an embedding-space MSE variant — hence *adapted*, not faithful (Appendix H). Anchor rows are the dedicated anchor runs of the control experiments, evaluated at the design test L=0.152; the fine-scan pool of Figure 1b (5 seeds, test L=0.150) gives 93.35 ± 0.71 for the same anchor. Within-table comparisons are same-protocol throughout.

Method	clean	test L=0.152	test L=0.200
AugMix w/o consistency (mixonly)	91.94 ± 0.40	8.15 ± 0.82	0.19 ± 0.08
AugMix (MSE consistency)	92.61 ± 0.28	36.13 ± 4.79	1.43 ± 0.58
MixTrain (mixture corruption training)	85.61 ± 0.34	84.09 ± 0.65	**52.41 ± 0.60**
Baseline (L=0)	92.34 ± 1.52	18.05 ± 1.22	0.48 ± 0.17
L_a (0.060)	92.79 ± 0.34	84.27 ± 1.02	12.50 ± 1.21
L_{opt} (0.152)	91.64 ± 0.99	**91.50 ± 0.88**	36.09 ± 3.45

AugMix implementation note. We reproduce AugMix's three components (K=3 chains, Dirichlet mixing, a consistency term) in the contrastive setting. The original logit-space JSD proved unstable here: at weight λ=12 (the classifier value) it collapsed training outright, and at λ=1 it destroyed already learned structure within a single epoch — even with consistency warmup and clean-branch detaching (probe logs in the supplement). We therefore report two working variants: mixture-only (the w/o-JSD row of the original paper's own ablation) and bounded embedding-space MSE consistency. The consistency term is worth +28.0 R@1 points at test L=0.152 (36.13 vs. 8.15), but both variants remain far below the fixed extreme anchor.

I CALIBRATION SENSITIVITY AND SCOPE RULES

θ validity: is the threshold schedule-invariant? For every schedule whose landmark was located, we measured the *joint* ring-match curve of its member operators and read the survival at the observed landmark: balanced 0.259 (at 0.152), N1 0.258 (at 0.156), N3 0.273 (at 0.158), N5 0.218 (at 0.1704), N8 0.266 (at 0.1314). The values cluster in $[0.22, 0.29]$ around a mean of $\approx$0.26 — close to, but systematically below, the product-axis calibration $\theta = 0.2961$, an average sub-multiplicative interaction shift of $\approx$0.03. Two conclusions: the critical-survival threshold is approximately schedule-invariant, and the predictor's misses stem from the product-independence assumption (and compensability),

not from the threshold being a magic number. N6's observed bracket peak shows joint survival 0.50 — far above the transition range — consistent with its response being a flat plateau rather than a true transition.

θ sensitivity. Perturbing the reference optimum within its bracket uncertainty ($L_{\text{opt}} \in [0.147, 0.157]$, i.e. $\theta \in [0.2494, 0.3334]$) yields prediction drift of ± 0.004 (N1), ± 0.012 (N5), and ± 0.035 (N3): sensitivity grows with extrapolation distance from the reference schedule. The N1/N5 hits are robust to calibration uncertainty; N3's miss is partly explained by N3 being the most sensitive schedule. Two further robustness checks across the same θ band: N7's negative prediction is unaffected (its product curve never falls below 0.52, far above the band's top), and V3's false positive stands (prediction ranges $[0.220, 0.241]$, inside the probed $[0.215, 0.245]$); N2's prediction ranges $[0.440, 0.482]$, so its false-positive verdict holds over $\theta \gtrsim 0.27$ and is censored at the band's low end, where the prediction drifts beyond the probed range. The direction of the θ drift is consistent with the reference-model analysis of Appendix C: recalibrating θ on the second reference moves predictions toward the observed optima of the located misses.

Interaction error, measured on the joint curves. The product rule is a first-order approximation. Measured joint ring-match curves deviate from the product of single-operator curves in both directions, with sign flips even within one pair across intensity: N3's pair (noise+grid) runs $+22\%$ at $L{=}0.02$ through $+25\%$ at 0.20 (mildly super-multiplicative), N5's pair (noise+downscale) swings from $+7\%$ at 0.02 to -36% at 0.16 and back to $+29\%$ at 0.20, and N8's pair (jpeg+grid) is within $\pm 6\%$ through the operating window but turns strongly super-multiplicative beyond it ($+57\%$ at 0.18, $+121\%$ at 0.20) — consistent with its undershoot. N6's pair (jpeg+downscale) tracks the product to within $\pm 5\%$ across the window. Quantifying the consequence for the hit: N5's joint curve crosses θ at 0.156 — 0.015 earlier than the product-based prediction — while the observed landmark sits at the prediction, so N5's exactness reflects a near-cancellation of interaction error (-0.015) and compensability ($+0.015$), not the smallness of either term. No simple frequency-overlap rule predicts the sign; we report the sign matrix as measured (supplementary CSVs).

Competition-zone rule. Early saturation of the effective knobs inside the operating window is one *sufficient* condition for the absence of a transition — not a necessary one. Balanced schedules vary both sides throughout (transition present); under 2:1 coupling the purifier side saturates at $L \geq 0.075$, and under 1:2 coupling the destructor side saturates at $L \geq 0.075$ — both yield flat plateaus, exactly as observed. At extreme coupling the picture extends: under corrected $k{=}3$ semantics (N2) the purifier side saturates at $L \approx 0.05$ while the uncapped destructor side keeps evolving, yet the measured response is still a flat, high plateau across the probed range $[0.445, 0.475]$ — the predicted cliff is a false positive. The saturation rule is derivable from the knob mapping alone and can be pre-registered for future schedules, but it does not explain every flat response.

N3 miss decomposition (corrected). The product rule's predicted crossing is 0.1133; the measured joint curve crosses θ at 0.14 — interaction error contributes 0.027 — and the observed optimum sits at 0.1733, a further 0.033 beyond the joint crossing, which we attribute to occlusion-compensability: anchor training learns to route around occluded patches. The direct evidence: under the same N3-pattern degradation, the N3-trained anchor recovers 91–95% of ground-truth rings where the clean reference model recovers 26% (measured on the reference; Appendix protocols). Full N3 curve (test $L{=}0.150$): 78.98 @ 0.0983, 88.27 @ 0.1133, 91.15 @ 0.1283, 91.75 @ 0.1433, 91.37 @ 0.1583, **91.98 @ 0.1733**.

Search-vs-measurement cost (amortized accounting). The honest ledger is: a one-time fixed cost — clean reference-model training ($\sim$1 GPU-hour), the 9-operator three-channel spectra ($\sim$1 GPU-hour), and the reference-schedule fine scan used to calibrate θ (a 5-seed valley scan at 0.002 resolution, $\sim$62.5 GPU-hours) — plus a marginal cost of $\sim$0.1 GPU-hour per new schedule (fine spectra for any new member operators; prediction itself is seconds). Scanning a new schedule directly costs $\sim$7.5 GPU-hours (a 3-anchor $\times$ 3-seed bracket). The total-cost breakeven against scanning is therefore $64.5/(7.5 - 0.1) \approx 9$ schedules within one coordinate system; at 10 schedules the cumulative saving is only $\approx 1.1\times$, growing toward the asymptotic marginal ratio of $75\times$ as the fixed cost amortizes. We state this plainly: measurement pays off for *programs* that design many

schedules in one coordinate system, not for a single deployment. All timings are on a single RTX 5090.

Table 8: Three deployment regimes on the reference schedule (fixed test L): naive is point-optimal, MixTrain dominates the mid range, a deep anchor wins only in the extreme tail.

Anchor	test 0.150	test 0.200	test 0.250	test 0.300
naive @ 0.150 (3 seeds)	**93.15 ± 0.57**	36.03 ± 1.98	7.17 ± 1.84	0.31 ± 0.14
MixTrain (3 seeds)	83.90 ± 2.01	**52.01 ± 3.68**	23.25 ± 2.35	5.50 ± 0.87
deep @ 0.450 (3 seeds)	59.13 ± 4.00	45.00 ± 3.42	**37.03 ± 3.45**	**26.28 ± 2.67**

N5 verification protocol (full). N5's bracket grid was generated by the pre-registered prediction script from the measured fine spectra *before* any training (the bracket is centered on the prediction by protocol design, recorded in `PREDICTIONS_N3N5.md` and `brackets.csv` with timestamps). Bracket results (test L=0.150, I2G R@1): 0.1554 → 90.74, 91.54, 89.48 (90.59 ± 1.04); **0.1704 → 91.74, 91.40, 91.14** (91.43 ± 0.30); 0.1854 → 90.56, 91.19, 90.71 (90.90 ± 0.33). The observed optimum coincides with the predicted point ($\Delta = 0.0000$); the two adjacent brackets are 0.53–0.84 points lower.

J AUDIO PILOT PROTOCOLS

Dataset and subset. FMA-small (Defferrard et al., 2017) (8,000 clips, 30 s, 8 balanced genres, CC-licensed). We use a stratified subset of 200 clips per genre (1,600 clips; first 10 s, 16 kHz mono) for the task and frequency channels, and a nested 50 per genre (400 clips) for the structure channel (pitch tracking cost). The subset is fixed by seed and released as `subset.csv`.

Reference model and calibration record. The task channel needs a competent frozen reference, the analog of the clean-trained molecular reference. Our first choice, the general-audio CLAP (`laion/clap-htsat-fused`), reached only 26% clean zero-shot genre accuracy (0% on Pop) — below our pre-set 40% competence criterion — because genre is weakly verbalizable in general audio captions; prompt engineering (6 candidates per genre, margin-ranked selection) lifted this only to ≈39%. We therefore switched to the music-specialized CLAP checkpoint (`music_audioset_epoch_15_esc_90.14.pt`, HTSAT-base), which reaches **51.5%** clean accuracy (vs. 12.5% chance) with a 6-prompt ensemble per genre. All calibration used clean clips only; no degraded data informed the reference or prompts.

Channels. *Task*: zero-shot genre accuracy (6-prompt ensemble text embeddings, cosine match). *Structure*: f0 agreement (`librosa.yin`, both-voiced frames, quarter-tone tolerance) and beat F-measure (70 ms window), degraded vs. clean — a tracker-estimated *pseudo-oracle*, weaker than the molecular ring-match ground truth, and reported as such. *Frequency*: high-frequency energy ratio (linear STFT magnitude, time-averaged, upper-half band over total), log2 change vs. clean — the direct analog of the image outer-half radial FFT ratio. (A literal 2-D-FFT-on-log-mel port was tried first and discarded: it is dominated by temporal modulation energy and does not isolate audio-frequency content.)

Operators and intensity coordinate. lowpass (8th-order Butterworth; cutoff 750–8000 Hz), mp3 codec (lameenc/miniaudio; 16–128 kbps; encoder delay corrected by cross-correlation alignment, without which frame-wise structure scores measure the *tracker*, not the signal), additive white noise (0–30 dB SNR), time dropout (100 ms blocks, 5–50%). Per-operator normalized $L \in [0, 1]$: $1 - c/8000$ (lowpass); $1 - \log_2 b / \log_2 128$ (mp3); $(30 - \mathrm{SNR})/30$ (noise); drop fraction (dropout). The classification window is $L \in [0.05, 0.5]$ with the paper's pre-registered thresholds.

Kill-criteria invocations. (i) Beat tracking returns no beats at extreme lowpass (750–1000 Hz): those cells are NaN and excluded from means. (ii) The mp3 structure channel before delay correction measured tracker sensitivity (flat ≈0.55 at all bitrates including 128 kbps); the channel was re-measured after alignment, and the bad data was discarded entirely.

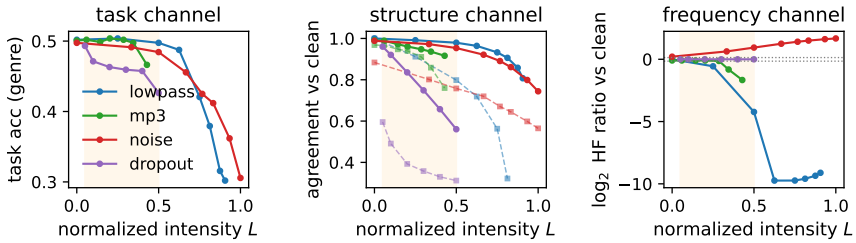

Figure 4: **Audio decay signatures** (FMA-small, music-CLAP reference). Task channel (left): lowpass is flat through the operating window then collapses threshold-style; noise decays gradually; mp3 is near-harmless. Structure channel (middle; f0 solid, beat dashed): time dropout erodes beat structure while task accuracy is nearly flat — a temporal-axis structure–task dissociation. Frequency channel (right): lowpass removes deeply, noise injects, mp3 removes mildly. Shaded band: classification window.

Table 9: Audio pilot full sweeps (means; task = genre accuracy, struct = agreement vs. clean, freq = $\log_2$ HF-ratio change vs. clean). Classification window $[0.05, 0.5]$.

Operator	knob	L	task	struct f0	struct beat	freq
lowpass	8000 Hz	0.000	0.502	1.000	1.000	0.00
	6000 Hz	0.250	0.504	0.991	0.912	−0.56
	4000 Hz	0.500	0.498	0.979	0.798	−4.23
	3000 Hz	0.625	0.488	0.964	0.719	−9.74
	2000 Hz	0.750	0.421	0.933	0.563	−9.74
	1500 Hz	0.812	0.379	0.906	0.322	−9.59
	1000 Hz	0.875	0.316	0.858	—	−9.35
	750 Hz	0.906	0.302	0.807	—	−9.12
mp3	128 kbps	0.000	0.501	0.991	0.968	−0.10
	96 kbps	0.059	0.502	0.989	0.967	−0.10
	64 kbps	0.143	0.499	0.972	0.947	−0.10
	48 kbps	0.202	0.504	0.961	0.940	−0.10
	32 kbps	0.286	0.502	0.948	0.911	−0.13
	24 kbps	0.345	0.498	0.936	0.859	−0.82
	16 kbps	0.429	0.466	0.916	0.760	−1.67
noise	30 dB	0.000	0.498	0.989	0.884	0.21
	20 dB	0.333	0.491	0.972	0.801	0.63
	15 dB	0.500	0.484	0.954	0.758	0.94
	10 dB	0.667	0.456	0.921	0.720	1.25
	7 dB	0.767	0.425	0.891	0.671	1.41
	5 dB	0.833	0.412	0.861	0.645	1.50
	2 dB	0.933	0.362	0.799	0.598	1.61
	0 dB	1.000	0.306	0.744	0.565	1.67
dropout	0.05	0.050	0.493	0.960	0.595	0.00
	0.10	0.100	0.471	0.920	0.491	0.00
	0.20	0.200	0.463	0.836	0.393	0.00
	0.30	0.300	0.459	0.749	0.358	0.00
	0.40	0.400	0.458	0.657	0.330	0.00
	0.50	0.500	0.426	0.561	0.312	−0.01

K SCAFFOLD-OOD EVALUATION (FULL TABLE)

Trainability margin probe (figure).

Table 10: Scaffold-OOD evaluation (held-out chemical scaffolds; 3 seeds, I2G R@1). The anchor advantage transfers with undiminished magnitude; ID values in parentheses.

Test L	Baseline (L=0)	L_a (0.060)	L_{opt} (0.152)
0.000	90.20 (92.34)	90.65 (92.79)	88.83 (91.64)
0.152	17.03 (18.05)	82.15 (84.27)	**88.49** (91.50)
0.200	0.40 (0.48)	11.35 (12.50)	**31.89** (36.09)

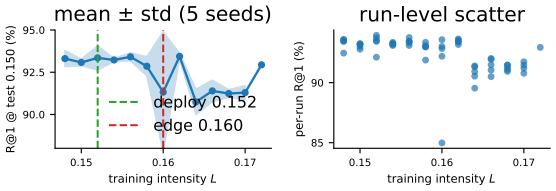

Figure 5: **Trainability margin at the plateau edge (balanced schedule, test L=0.150, 5 seeds per point).** Left: mean ± std across the plateau edge; variance is tight on the plateau and spikes at L=0.160 (one seed lands low). Right: run-level scatter. The deployable anchor retreats from the edge to the plateau interior (0.152–0.156).