THE BOUND DOMINATES THE BUDGET: CERTIFIED EXPLANATIONS FOR ADDITIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

How stable a minimal explanation of an additive model is depends far more on *which sound bound certifies it* than on how many features it names. A greedy minimal explainer stops the instant its error budget is met, spending 85.2–98.7% of the tolerance ε under its own unsigned acceptance test; certified against that test the radius is near zero, and buying 0.8 extra features moves it from 0.009 to $\mathbf{0.29}$ on `heart` at $r_{\max}$=0.5. But that test is not the predicate the explanation makes. The displayed surrogate's error is a sum of *signed* one-dimensional terms whose worst case over the ball is attained in closed form for main effects (Prop. 4); certified against it, the *same* minimal explanations have spent 32–38% and already reach 0.34 on `heart`, and the knob buys a further $\mathbf{1.2}\times$, not 31.7 (at the wider $r_{\max}$=2.0 error table, 0.45 and $\mathbf{1.5}\times$, not 14.9). The free change is worth more than the one we sell.

Measuring this needs a certificate rather than a sampled sensitivity score, and additive structure makes one cheap: for $\hat{y} = \sum_i f_i(x_i)$ the L_∞ worst-case output change decomposes *exactly* across features, so a sound per-input radius is d one-dimensional interval problems. We make that practical for *continuous neural* shape functions — dense-grid extrema padded by a *certified* analytic operator-norm Lipschitz bound — and pair it with a per-feature worst-case error table, so the certified object is the *displayed surrogate*: a subset *and* its functional forms.

The same holds one level up: what a backbone leaves after minimality differs up to $11\times$ across model classes unsigned and $\mathbf{1.3}$–$\mathbf{1.4}\times$ signed — both off the censoring ceiling, and only on `heart` is there an order-of-magnitude gap for the bound to close — so the choice of backbone is worth an order of magnitude less than the choice of bound — though which backbone leads is not stable, and EBM keeps `heart`. We instantiate on VIB-NAM, a bottleneck+bypass additive net, and are explicit that the bottleneck carries none of the results. Scope, costs and the failed tests are stated in §1; a provenance gate recomputes every headline number from the shipped artifacts.

1 INTRODUCTION

A minimal explanation — the smallest feature subset that reproduces a prediction within tolerance — is the object explainable ML has spent a decade learning to compute. We set out to report that minimality makes it fragile, and ended up reporting something else.

The mechanism is real: a greedy explainer accepts the first subset meeting its budget ε, spending 85.2–98.7% of it at the unperturbed point, so the slack a perturbation needs is what minimality consumed. But that budget is the explainer's own acceptance test, $\sum_{i \notin S} |f_i(x_i)|$, and it does not govern robustness: the displayed surrogate's error is a sum of *signed* terms. Certified against *that* — the predicate the explanation makes — the same minimal explanations spend 32–38% and 92–99% of them certify $r \geq 0.1$ (App. A49). **Which sound bound you certify with matters more than how many features you buy**: on `heart` the bound is free and worth $37.8\times$, the features cost 0.8 and are worth $1.2\times$ ($19.1\times$ and $1.5\times$ off the censoring ceiling).

The certificate scales and the explanation does not — that is this paper's scope. Both levers are real and we quantify both. Buying features means selecting against $\varepsilon_{\text{sel}} < \varepsilon$ and certifying against ε; the leftover budget is what a perturbation may occupy. On `heart` 0.8 extra features move the radius

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.

from 0.009 to 0.29 unsigned. We report the price, not only the benefit. The stabilised explanation names 67–91% of the features at d=13, and within a synthetic sweep holding the informative count at 6 the fraction grows with d: $|S|/d$ is 58%, 65%, 73% at d=20, 50, 100 (three seeds, small capacity, in-sample; not one sequence with the d=13 point). Under the unsigned test the minimal explanation certifies nothing at $d \geq 50$ and 20% of `spambase` (d=57); under arm D it certifies every instance, so the d-collapse too is a statement about the bound (App. A40). The cause is the criterion, not the search — a residual budget summed over *excluded* features accumulates past ε once there are enough of them. We test that directly under the headline protocol (five seeds): the *exact MIP optimum* picks the same $|S|$ as the greedy and, under arm D, certifies 0.97–$0.98\times$ its radius, below it on every dataset; under arm A it stays near zero on `heart` (0.014 against 0.009). Even the best of *all* minimum-cost explanations certifies $r \geq 0.1$ unsigned on 2–50% of instances, the greedy under arm D on 92–99% (App. A41).

Why this needed a certificate. Prior work measures explanation sensitivity by sampling perturbations (Yeh et al., 2019; Ghorbani et al., 2019). A sampled rate cannot distinguish "did not move on these draws" from "cannot move"; a frontier needs the second, and additive structure makes it cheap: the L_∞ worst case decomposes *exactly* across features, so a sound per-input radius is d one-dimensional interval problems rather than a relaxation over a network. We make that work for *continuous neural* shape functions and pair it with a per-feature error table, so what is certified is the *displayed surrogate* — a subset *and* its forms.

What we instantiate them on, and what that does not contribute. Our testbed VIB-NAM is a differentiable additive backbone whose shape function is $g_i^{\mathrm{nn}}(x_i; \beta_i) + w_i x_i + b_i$: a variational information-bottleneck path with swept per-feature compression β_i plus an unconstrained linear bypass, which degrades to a linear—not constant—effect under compression (Thm 1; the mechanism mirrors a penalized spline shrinking to its penalty's null space (Wood, 2017)). The bottleneck is a diagnostic, not the source of the results: it adds no accuracy, its certified-radius gain does not survive honest β selection, and the β–MI bound one might expect is false (App. A29). A reader who wants only the frontier can substitute any differentiable additive model; we run the certificates on two others below.

Contributions.

- **At minimality, the certification bound dominates the selection budget (§5.3, App. A49)** — the paper's main claim. Against the explainer's own unsigned acceptance test the minimal explanation consumes 85.2–98.7% of ε and certifies almost nothing; against the signed residual it actually displays it consumes 32–38% and certifies 0.34 on `heart` (0.45 at $r_{\max}$=2.0). Buying 0.8 features adds $1.2\times$ more ($1.5\times$ there), against the $31.7\times$ the unsigned bound credits it with. Both are sound; one is free. We report the cost as $|S|/d$ (67–91%), not as a count: $|S|$ is fixed by the selection condition and grows under *any* ball, while under a third of the radius collapse is the L_∞ threat model — over an L_2 budget the same explanations keep $\mathbf{1.7}$–$\mathbf{2.5}\times$ more radius off the ceiling (App. A45, arm A both sides).
- **Two certificates that make the frontier measurable (§4, §3.2)**. (i) A *sound* per-input L_∞ radius from the exact additive interval decomposition, extended to continuous neural shape functions by a certified analytic operator-norm Lipschitz padding (Thm 3); geometric-mean $\mathbf{1.87}\times$ (5 seeds, dataset-clustered $[1.48, 2.35]$) the radius a CROWN-verified protocol-matched MLP admits as selected; a validation-only rule finds a model at least as accurate in 15 of 45 cells, and all 15 certify wider (App. A39). (ii) VIB-X, a greedy minimal explainer shipping a *solver-free per-instance optimality bound* (LP relaxation of a stated conservative surrogate, valid for all d), within 0.9% of the MIP optimum at 11–$21\times$ lower latency. Composed, they certify the displayed explanation itself. Both apply to any additive model, which turns the certificate into an instrument: across three backbones what minimality appears to leave is set by the bound and not by the class, up to $11\times$ apart unsigned and $\mathbf{1.3}$–$\mathbf{1.4}\times$ signed on the same models — on `heart`; on two of the four datasets the signed spread is the wider one and there is nothing to collapse (§5.3, App. A46).
- **The guarantee is not vacuous, and we bound what it is worth (§5.3)**. Inside the certified ball the exact additive Shapley value flips its top-k set on 29% of perturbations $[25, 34]$ and KernelSHAP on 30% $[24, 36]$ (five seeds): exact Shapley samples nothing, so this fragility does not need sampling. The paired gap, 0.8 points $[-6.1, 7.6]$ (dataset-clustered), is too

wide to bound sampling's share or to call the two equal. On set stability our own re-selection (0.62) is *worse* than exact Shapley (0.71) on every seed (-0.09 $[-0.12, -0.05]$), and we report it.

What we do not claim. The bottleneck that names our testbed contributes none of the above: it adds no accuracy (mean $\Delta{=}-0.005$), its certified-radius gain fails a preregistered test under honest β selection, the multimodal end-to-end advantage we measured (reported in full in earlier revisions; runners and artifacts ship with the code) is a property of differentiable *additivity* (plain-NAM and NODE-GAM heads match it) and disappears against an interaction-aware GA2M baseline when the encoder is frozen, and the β–conditional-MI bound one might expect is false (App. A29). On nested-CV accuracy VIB-NAM sits in the top Nemenyi cluster without beating EBM; we claim no SOTA accuracy. The certificates also now cover the pairwise terms exactly as well (Prop. 1), and certifying that configuration costs radius rather than accuracy: geometric mean $\mathbf{0.44}\times$ the main-effects radius for $+0.022$ mean score (§4). We state these here rather than in a limitations paragraph because they define what the paper is about.

2 Related Work

Neural Additive Models and GAM Variants. NAMs (Agarwal et al., 2021) learn shape functions with ExU units; NODE-GAM (Chang et al., 2022) uses differentiable trees, SIAN (Enouen & Liu, 2022) interaction detection with sparse selection, GP-NAM (Zhang et al., 2024a) GP priors, KAN (Liu et al., 2025) learnable spline activations, NBM (Radenovic et al., 2022) a shared basis decomposition that scales GAMs by an order of magnitude in d, and EBM (Nori et al., 2019) boosted trees with pairwise interactions. VIB-NAM differs by coupling per-feature information bottlenecks with a linear bypass for data-driven capacity allocation. We note NBM specifically because our explanations degrade with d (App. A40) and a shared-basis head retains the per-feature one-dimensional structure both certificates need.

Information Bottleneck and Neural Processes. The IB principle (Tishby et al., 1999) trades compression against prediction and Alemi et al. (2017) gave the reparameterizable variational objective; VIB-NAM *parameterises* VIB per feature, though no experiment here exercises that freedom — all of them fix one β across features, so what we run is a global IB. Neural Processes (Garnelo et al., 2018; Kim et al., 2019) encode context sets into latent variables; we place an Attentive NP inside each shape function for native epistemic+aleatoric uncertainty without ensembles (evidentially in Schleibaum et al., 2026).

Certified and Minimal Explanations. Abductive explanations (Ignatiev et al., 2019; Bassan & Katz, 2023) and their distance-restricted (Izza et al., 2024a), inflated (Izza et al., 2024b) and probabilistic (Wäldchen et al., 2021; Izza et al., 2023) forms give sufficient reasons and bound how far, how widely and how surely they hold; Huang & Marques-Silva (2023) compute them with a number of adversarial-robustness queries linear in d, and Bassan et al. (2026) compute them exactly for neural additive models by a solver-dependent search whose complexity Bassan et al. (2025) chart across GAM classes. Anchors (Ribeiro et al., 2018) and SHAP (Lundberg & Lee, 2017) are local and post-hoc, without optimality guarantees. VIB-X sits between an exact solver — NP-hard in general, though Bassan et al. (2026) reach the optimum for neural additive models in $O(\log d)$ verification queries — and an unbounded heuristic: a millisecond greedy explanation *with* a solver-free LP optimality bound valid for every d (§5.2). Robustness-side, §4 generalizes the stump-ensemble separability of Andriushchenko & Hein (2019) to continuous neural shape functions, against the loose relaxations (Xu et al., 2020) unstructured networks require.

3 Method

3.1 The Testbed, in One Paragraph

Everything below needs some differentiable additive model; ours is VIB-NAM, which decomposes $\hat{y} = b + \sum_i f_i(x_i) + \sum_{(i,j)\in\mathcal{E}} h_{ij}(x_i, x_j)$ with

$$f_i(x_i) = g_i^{\mathrm{nn}}(x_i; \beta_i) + w_i x_i + b_i, \tag{1}$$

an unconstrained linear bypass plus a per-feature variational information bottleneck of strength β_i, trained in four phases (warmup, anneal, residual-screened pairwise interactions, joint fine-tune). β *trains* under the default schedule and is frozen on a grid only where a specific value is the object of study (§4's certified-radius sweep). Since $D_{\mathrm{KL}}[\cdot \,\|\, \cdot] \geq 0$, $\partial\mathcal{L}/\partial\beta_i \geq 0$, so that trained β collapses toward its floor and carries no data-driven meaning — which is why no result here rests on it. The architecture, the schedule, the interaction screen, the graceful-degradation theorem (Thm. 1) and the scope of what β constrains are all in App. A3, because none of them is load-bearing for the results: §4 and §3.2 apply to any additive model, and a reader who wants only the frontier can substitute their own backbone. What the bottleneck does *not* buy is measured in §5.1.

3.2 VIB-X: Approximate Minimal Explanations

Given a trained VIB-NAM and test point $\mathbf{x}$, VIB-X finds a small set $S \subseteq [d]$ and simple approximating functions $\{\tilde{f}_i\}_{i\in S} \in \{\mathrm{const}, \mathrm{linear}, \mathrm{quad}, \mathrm{full}\}$ whose residual satisfies $|\hat{y} - b - \sum_{i\in S} \tilde{f}_i(x_i)| \leq \varepsilon$. **Certified program (disclosure).** All guarantees (greedy, MIP, and the LP certificate) are stated for a *conservative surrogate*: using offline per-feature worst-case errors $e_i^{\mathcal{C}} = \max_x |f_i(x) - \tilde{f}_i^{\mathcal{C}}(x)|$ we enforce $\sum_{i\notin S} |f_i(x_i)| + \sum_{i\in S} e_i^{\mathcal{C}_i} \leq \varepsilon$, which upper-bounds the displayed residual. Every certified explanation is therefore feasible for the displayed program, but the optimality gap is with respect to the surrogate — sign cancellation could admit cheaper explanations under the displayed residual (derivation in App. A49). VIB-X adds features in descending $|f_i(x_i)|$ order until the residual is $\leq \varepsilon$, picking the simplest class per feature. The program is a multiple-choice covering knapsack (App. A36); the greedy has no a priori ratio, so we certify sub-optimality *per instance* via its LP relaxation. Empirically the mean ratio is ≤ 1.009, with 51–100% of predictions matching the exact MIP oracle on seven datasets ($d \in \{8, \ldots, 13\}$, §5.2). The interaction-aware extension is left to future work.

4 Sound Robustness Certification

The additive structure of VIB-NAM buys something black boxes obtain only through looser relaxations: the reachable output set under an L_∞ perturbation *decomposes exactly* across features. For $f(\mathbf{x}) = \sum_i f_i(x_i)$, writing $U_i = \max_{|\delta_i|\leq\varepsilon} f_i(x_i{+}\delta_i) - f_i(x_i)$ and $L_i = \min_{|\delta_i|\leq\varepsilon} f_i(x_i{+}\delta_i) - f_i(x_i)$,

$$\Big\{ f(\mathbf{x}{+}\boldsymbol{\delta}) - f(\mathbf{x}) : \|\boldsymbol{\delta}\|_\infty \leq \varepsilon \Big\} = \Big[\textstyle\sum_i L_i,\ \sum_i U_i \Big],$$
$$\max_{\|\boldsymbol{\delta}\|_\infty \leq \varepsilon} \big| f(\mathbf{x}{+}\boldsymbol{\delta}) - f(\mathbf{x}) \big| = \max\Big(\textstyle\sum_i U_i,\ -\sum_i L_i \Big), \tag{2}$$

because the ball's coordinates vary independently and each continuous f_i fills $[L_i, U_i]$. The *signed* form matters: the naive $\sum_i \max(U_i, -L_i)$ ignores cancellation and can overstate the worst case. Each U_i, L_i is one-dimensional, so the interval costs $O(dG)$ per input — milliseconds. **The certified radius** is the largest ε keeping the prediction usable: worst-case $|\Delta f| \leq \tau$ (regression) or no logit sign flip (classification), ε in z-scored feature units and $\tau{=}0.5$ in standardised target units, the two predicates reported apart in App. A39. This generalizes the stump-ensemble observation of Andriushchenko & Hein (2019) to continuous neural/spline shape functions; there the grid evaluation is sound up to a *certified* padding $L_i^{\mathrm{lip}} h$, with L_i^{lip} an analytic operator-norm Lipschitz bound — *not* an observed gradient estimate — and h the grid step (Theorem 3, Appendix A21; sound in exact arithmetic — the float32 run carries no rounding budget). Pairwise interactions take the same per-block treatment, and for the lookup-table shapes we use *each block* is exact: on a cell the interpolant is bilinear, hence affine along each axis, so over an L_∞ box it is extremised at the corners of the box clipped to each cell — a small closed-form candidate set, no padding (Prop. 1). **The model's interval is then sound but not attained**: f_i and each h_{ij} containing i share x_i yet are maximised separately, so the sum relaxes — to 1.3–1.8× a lower bound on the true reachable width here, against 1.00× for the main-effects model, whose blocks are functions of distinct coordinates (App. A23).

So the certified model can be the deployed model, and the trade is now priced. Certifying the interaction-bearing model instead costs radius on all four datasets (0.18–0.87×, geometric mean $\mathbf{0.44}\times$) and buys $+0.022$ mean test score that we cannot resolve (95% CI $[-0.032, +0.075]$, four

clusters, 40 test points per seed). **It costs radius consistently and buys accuracy we cannot measure**, and part of that cost is our own relaxation.

Glass-box certifies $1.87\times$ wider radii than a verified MLP. On 9 UCI datasets ($\leq$ 100 test points each, five seeds, 5 frozen β values) we compare VIB-NAM's certified radius to a regularized MLP baseline (matched training protocol; two hidden layers against VIB-NAM's three, and carrying $\sim$30–100$\times$ fewer parameters — both gaps are closed and separated below) certified with `auto_LiRPA` CROWN bounds (Xu et al., 2020). The asymmetry is the point: CROWN yields a *lower bound* (the best a deployer can guarantee for the MLP), whereas VIB-NAM's radius is sound with an explicit, certified slack: dense-grid intervals ($G{=}10^6$ knots/feature) padded by an analytic operator-norm Lipschitz bound (Theorem 3; total certified padding 0.004–0.101 across all (dataset, β, seed) cells, at most 20.3% of $\tau{=}0.5$, stored in each artifact). The comparison contrasts what each *deployer can prove*, not oracle robustness. Under our *preregistered validation-selection protocol* (β chosen per dataset among values whose validation clean score is within 1% relative of the best, by mean certified radius on 50 held-out *validation* points; test points touched only after selection) the certified radius is geometric-mean $\mathbf{1.87}\times$ the MLP's CROWN bound over **five seeds** (42–46; per-seed 1.81, 1.88, 1.87, 1.92, 1.86; SD(log) = 0.022, 95% t-CI $[1.82, 1.92]$ *between seeds*). Seeds are the wrong unit for generalisation: clustering by *dataset* gives $[1.48, 2.35]$, nine times wider, and that is the interval a reader should carry. We use the cluster-t for *both* clustered claims in this paper rather than choosing per claim. Both it and the pairs bootstrap are in the artifact (`results/robust_cert/cluster_intervals.json`) (App. A39). The *best-β test envelope* — an explicitly oracle upper bound, never our headline — is $2.04\times$ $[1.98, 2.11]$ over five seeds (Figure 11a draws seed 42, where it is $2.01\times$; the per-β radii it envelopes are the `vibnam_beta*` entries of the certification artifacts named below). At $\varepsilon{=}0.25$ the MLP certifies $\approx 0\%$ on five of six regression tasks and VIB-NAM 0–100% over the five seeds — `california` 0% on every seed, our one loss, against `diabetes_reg` at 95–100%.

How far five seeds get us, and three ways it could be unfair. Seed-to-seed dispersion is small (2.2% on the log scale) and seed 42 is the *least* favourable of the five. Three objections to comparing a near-exact certificate against a CROWN *relaxation* each have a control in App. A39, including the one that goes against us: against sound α-CROWN the headline barely moves ($\mathbf{1.78}\times$), against a parameter-matched MLP it *grows* to $2.35\times$, and in the depth-matched cell it is $\mathbf{1.86}\times$ — while on `california` the MLP provably certifies wider, which is a real reversal and not a relaxation artifact.

Two results that do not go our way. The preregistered test (`C5` in the registration, G3 here) — compression should lift the certified radius — **fails**: validation-selected β picks $\beta{=}0$ on 2–5 of 9 datasets depending on the seed, so β is a diagnostic, not a knob a deployer can bank on (App. A55). And although a PGD study corroborates the certificate within the differentiable class ($7/9$, one-sided $p{=}0.027$, persisting under FGSM, CW-L_2 and Folktables shift (Ding et al., 2021)), EBM scores higher under the random attack its lack of gradients limits it to — a weaker attack, so we claim no ordering against EBM — and it is a win-rate over a losing family: at PGD $\varepsilon{=}0.20$ our attacked R^2 is at or barely above the mean predictor on four of six regression sets (`california` -0.257, `wine` 0.015, `abalone` 0.055, `concrete` 0.089 — only `california` is actually below it) where EBM holds 0.245 and 0.273; we claim only that within the differentiable class VIB-NAM is the most robust option *and* uniquely admits the exact certificate (Apps. A56, A57).

5 Tabular Accuracy Benchmark

We evaluate VIB-NAM on nine benchmark datasets spanning regression and classification tasks under a nested 5-fold CV protocol (seed 42; outer 5-fold KFold; inner 20% validation split used for early stopping, phase selection, and boost-round acceptance; held-out outer fold used *only* for the reported score; same folds, unequal refit budgets, App. A9). Ablation, calibration, and VIB-X studies use 3-fold CV. Metrics are R^2 (regression) and AUROC (classification). We compare against **NAM** (Agarwal et al., 2021), **NODE-GAM** (Chang et al., 2022), **EBM** (Nori et al., 2019), **XGBoost** (Chen & Guestrin, 2016), and **Linear** (logistic or ridge regression). EBM is at library defaults, NODE-GAM is not (App. A9); XGBoost uses 300 estimators, depth 6, lr 0.1, subsample 0.8. VIB-NAM uses warmup / anneal / interaction / fine-tune epochs of $(60, 120, 150, 60)$ with Adam. Baseline settings

and hardware are in Appendix A4; library versions are recorded per artifact rather than in the paper (`results/ENVIRONMENT.json` and every artifact's `_metadata`).

5.1 Main Results

Accuracy is a sanity check, not a claim: the certificates and the frontier are properties of additive structure, so this section need only establish that the testbed is not a strawman. Table 24 (App. A37) reports nested-CV performance on nine datasets. Average ranks put EBM first (2.89) and VIB-NAM-Boost-Ens second (3.28), with base VIB-NAM at 4.39. VIB-NAM-Boost-Ens is VIB-NAM plus shrunk per-feature boosting rounds (App. A37); it stays additive, so everything certified here applies to it unchanged. VIB-NAM-Boost-Ens is not significantly different from EBM ($\Delta{=}-0.006$, $p{=}0.43$; no equivalence margin was set, so not a tie) and XGBoost ($p{=}1.0$); its uncorrected wins over NODE-GAM and Linear ($p{=}0.027$, 0.020) *do not survive Holm* over the seven-comparison family (adjusted 0.14, 0.12), and Friedman is $p{=}0.058$ — not a strict rejection, so the Nemenyi diagram (Fig. 2) is reported as descriptive rather than as a post-hoc test. The honest summary is *competitive with EBM/XGBoost, beating neither*. Full table, per-fold scores and extended datasets: Apps. A37, A6.

Cost and deep-tabular baselines. Mean training time per fold: VIB-NAM 58s, VIB-NAM-Boost-Ens 214s, NAM 37s, EBM 47s, XGBoost $<$ 1s — the ensemble is about $4.6\times$ slower than EBM; base VIB-NAM is faster than the ensemble, not than EBM or NAM (App. A14). **TabM** (Gorishniy et al., 2025), **SAINT** (Somepalli et al., 2021) and **TabTransformer** (Huang et al., 2020) are in App. A7; they subsample to 10^4 rows untuned, on other test sets without target normalisation, so that table supports no ordering and we draw none.

Component isolation (matched capacity). At identical $(256, 256, 128)$ capacity the IB+bypass adds no accuracy over a matched plain NAM (mean $\Delta_1 = -0.005$); the interaction phase carries the gain (mean $+0.042$, confounded with its extra training), and VIB-NAM-NP's intervals over-cover ($\text{PICP}_{95} \approx 1.0$) (Apps. A8, A10).

5.2 VIB-X Optimality Analysis

Against exact MIP solutions on the seven $d \leq 13$ datasets (100 sampled predictions each) greedy is near-optimal — mean cost ratio ≤ 1.009 (cost: one per feature plus a tenth per form level, App. A36), optimal on 51–100% of predictions (Table 23). These figures need the error table built at the certifier's grid density: a coarser table inflates every e_i^C until the choice set collapses to `full` and exactness becomes spuriously high on a degenerate program (App. A14).

Per-instance optimality certificates (no solver). The MIP comparison is capped at $d{\leq}13$ and a sampled subset; a solver-free certificate closes both gaps. The program is a multiple-choice knapsack whose LP relaxation has a closed-form optimum: we take it with the sorted convex-hull greedy of Sinha & Zoltners (1979), $O(dL \log dL)$, which Dyer (1984) and Zemel (1984) improve to $O(dL)$ by prune-and-search — at this size we do not need their linear-time form; since $\text{LP} \leq \text{OPT} \leq \text{cost}_{\text{greedy}}$, every explanation ships with a certified ratio $\rho(\mathbf{x}) = \text{cost}_{\text{greedy}}/\text{LP}$ bounding its sub-optimality *w.r.t. the surrogate program of §3.2*, for all d and every test point. It is sound on $443/443$ MIP-verified points, with the error table tabulated over a range that provably *covers* every certified point, so no query is answered by extrapolation. Median certified gap 1.5–8.9% against a mean greedy gap of 0.31% (81% exact), so the residual is largely LP integrality rather than greedy error.

5.3 Certified-Stable Explanations: What the Two Certificates Buy Together

What a practitioner asks of an explanation is whether it still holds after a small, bounded change of input. Post-hoc explanations are famously fragile (Ghorbani et al., 2019; Alvarez-Melis & Jaakkola, 2018) and can be adversarially manipulated (Slack et al., 2020); Yeh et al. (2019) make sensitivity a *measured* quantity and our top-k flip rate below is its set-valued analogue. Measuring sensitivity and *certifying* its absence are different deliverables: the first says how often an explanation moved on the sample drawn, the second that it cannot move inside a stated radius.

What is and is not new here. Formal XAI already certifies explanations (Ignatiev et al., 2019; La Malfa et al., 2021; Wu et al., 2023), and the distance-restricted AXp of Izza et al. (2024a) is a sound per-instance radius on which an explanation stays a valid sufficient reason: "no prior work certifies an explanation under bounded perturbation" would be false and we do not say it. What is ours is the *displayed surrogate* — a subset *and* its forms — shown to approximate f within ε at every point of the ball, the named features moving too. Bassan et al. (2026) already cover regression, but by holding x_S fixed and bounding f's own deviation from $f(x)$ as the complement moves; ours is a fidelity guarantee on a simpler function, in closed form. At our ε their predicate is met by the *empty* explanation at $r \geq 0.1$ on 87–100% of instances; ours also requires the displayed function to match the prediction's level (App. A38). That is a narrow distinction and we price it: at the `full` level our radius *is* a distance-restricted sufficiency radius, and restoring every form to `full` changes it by 1.02–1.12× under the bound we argue for. App. A38 works the comparison against Bassan et al. (2026), who reach the cardinal optimum on this model class in $O(\log d)$ verification queries; none of it supports our thesis.

For an additive model the computation is exact. **Scope**: certifying a VIB-X explanation — a subset *plus* its forms — needs the error table, since $e_i^{\mathcal{C}_i}$ enters the condition. Certificate 1 alone certifies only that the top-k *set* cannot change, and is not vacuous on its own (0.111–0.160 mean radius, positive on 96–100% of instances); the two certify different predicates and we do not read one against the other (App. A54).

The construction. VIB-X accepts S when $\sum_{i \notin S} |f_i(x_i)| + \sum_{i \in S} e_i^{\mathcal{C}_i} \leq \varepsilon$. The second sum is a property of the *model*, so no perturbation moves it; the first is a sum of *one-dimensional* functions, so under $\|\delta\|_\infty \leq r$ each term is bounded exactly by $M_i(r) = \max(|f_i + U_i(r)|, |f_i + L_i(r)|)$ — the padded interval §4 already computes. The condition is monotone in r, so the largest certified radius follows by bisection, in milliseconds, inheriting the analytic-Lipschitz soundness. We certify *validity*; that the greedy would *rediscover* the same S is stronger, and App. A53 certifies that too — on a ball 0.005–0.082× as large, which the knob does not enlarge.

Two ways to bound that first sum, and the gap between them is this section. Term by term — VIB-X's own acceptance test (**arm A**) — discards every sign, charging the budget as if all excluded features moved against us at once. The residual the explanation *displays* is signed, and over a box its worst case separates coordinatewise and is *attained* (**arm D**, Prop. 4): same models, same selections, one bound declining to throw the cancellation away.

Under its own acceptance test, minimality makes the explanation fragile; under the predicate it displays, it does not. The greedy stops the instant the budget is met, spending 85.2–98.7% of ε over five seeds under arm A (Fig. 1b, hollow markers), so the radius certified there is near zero. Under arm D (Fig. 1, filled markers) the *same* minimal explanations have spent 32–38%, and on `heart` 92% of them certify $r \geq 0.1$ rather than 0%, already reaching 0.34 (0.45 at $r_{\max}$=2.0), more than the 0.29 below buys with 0.8 features; the knob adds 1.2×, not 31.7× (there 1.5×, not 14.9×).

This is a claim about certifying, not about selecting. Selecting against the signed residual too (arm C) certifies $r \geq 0.1$ on only 40/46/52/12% of instances, worse than arm A on `abalone`: what the bound buys, it buys on the selection the explainer already makes. What is displaced is the *explainer's own* acceptance test, not a bound the literature recommends (App. A38).

Nor is it the additive gauge. Shapes are identified only up to constants, and the unsigned test is not invariant to them; centring every shape on its training mean changes most minimal subsets yet leaves the bound worth 2.3–33.1× and the knob 1.21–1.32× (App. A50).

The budget is a knob, and it is the smaller lever. Selecting against a tighter $\varepsilon_{\text{sel}} < \varepsilon$ while certifying against ε turns this into a knob: on `heart`, ε_{sel}=0.35 costs 0.8 extra features (7.9→8.8) and moves the mean certified radius from 0.009 to **0.29** at $r_{\max}$=0.5 (2.5–31.7× over the four datasets; the same two points are 0.024 and 0.352 at the $r_{\max}$=2.0 Fig. 1a is drawn at), with certifiable instances going from 90% to **100**%.

Raising $r_{\max}$ is not a decensoring; `--r_max` also sets the error table's margin, so it inflates every $e_i^{\mathcal{C}}$. Both minimal columns are uncensored (0% of `heart` instances at either ceiling), so all of 0.009→0.024 is the margin — decensoring alone would move the ratio *up*. We write "off the ceiling" below only as a name for the $r_{\max}$=2.0 run, no frontier ratio we report is invariant to $r_{\max}$, and we report both (App. A49).

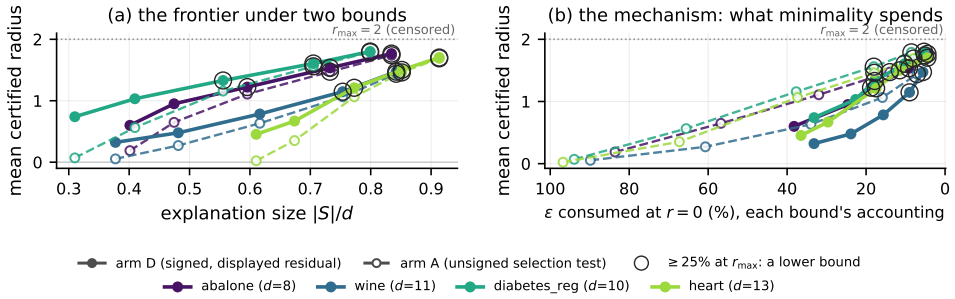

Figure 1: **The frontier, and the bound it is measured with.** Clean protocol, five seeds, the *same* selections certified two ways: **hollow, dashed** is the explainer's unsigned acceptance test (arm A), **filled, solid** the signed residual it displays (arm D, exact by Prop. 4, main-effects premise). The vertical gap between a dataset's curves is what the bound is worth; movement along a curve is what the budget is worth, and it is the smaller. **(a)** radius vs. $|S|/d$: the minimal explanation is each curve's bottom-left end and tightening ε_{sel} walks up and right toward $|S|/d \to 1$. **(b)** the mechanism: radius against the budget consumed at $r{=}0$ under each bound's own accounting — arm A spends nearly all of ε and keeps no radius; arm D has spent a third and keeps most. Both arms are read from *one* run at $r_{\max}{=}2.0$, so the two budgets the body discusses are drawn essentially uncensored (0–2.5% of arm-A instances at the ceiling at $\varepsilon_{\text{sel}}{=}\varepsilon$ and 0.35; at $r_{\max}{=}0.5$ they are 46–91% censored by $\varepsilon_{\text{sel}}{=}0.20$ and the curves meet on the ceiling line — an artifact, not a finding). Dotted line: $r_{\max}{=}2.0$; **a ringed marker** is a mean with $\geq 25\%$ of its instances there, i.e. a lower bound — 18 of the 40 points: both tightest budgets entire, plus arm D at $\varepsilon_{\text{sel}}{=}0.20$ on `abalone` 25%, `heart` 27% and `diabetes_reg` 28% (`wine`'s arm A at $\varepsilon_{\text{sel}}{=}0.1$ is just under, at 24%). `.../armd_uncensored_seed42.json` and seeds 43–46; plotted values in `paper/figures/frontier.json`, regenerated by the figure gate.

Those numbers are measured with *both* disclosed protocol defects repaired at once — benchmark capacity and a genuine validation split (App. A48). **What ε is a fraction of**: at $\varepsilon{=}0.5$ the tolerance is 0.24 of `heart`'s own output spread and 0.97 of `wine`'s, so a radius on `wine` sits inside a tolerance nearly as wide as the prediction it approximates. Each repair alone moves the number the other way. This *stability–simplicity frontier* (Fig. 1; one patient of it, with both budget bars, in Fig. 12, App. A49) is smooth and cheap, measured at $G{=}10^6$ to match §4. A size–fidelity trade is not new (Macdonald et al., 2019; Yeh et al., 2019); the certificate on the displayed explanation is.

It transfers, and the "headroom" between backbones is mostly the bound too. Both certificates run on EBM, whose piecewise-constant effects admit the interval *exactly*, with no padding: in the small-capacity in-sample arm its minimal explanations spend 79–91% of the budget against our 87–98%, and tightening buys 1.9–4.7× radius for 0.7–1.1 extra features — the same trade. What differs is what the spent budget *leaves*, and unsigned that looks like a property of the backbone: every additive backbone we certify spends 77–96% of ε on its minimal explanation, and the spread is 1.2/1.1/2.9× on `abalone`/`wine`/`diabetes_reg` but 11× on `heart`, where the unsigned criterion certifies 0.262 for EBM, 0.059 for a plain NAM and 0.024 for ours — *under arm A*, EBM ahead 3/4 — a gap that reads as advice to fit a histogram. **It is not.** The signed bound localises the very surrogate-error term that gap was made of, and on *those* models arm D leaves 1.27–1.41×, EBM keeping `heart` (0.514 against our 0.454). The win count is not readable — 3/4 uncensored, 2/4 censored, 2/5–5/5 per seed — so we *withdraw* any claim that the ordering reverses. The collapse is the finding; the ordering is not. The backbone is worth at most 1.41× between seed means on the clean protocol, 2.56× within a seed (Apps. A42, A46).

A third axis, and the same artifact. The greedy also downgrades each selected feature to the cheapest form that fits. Holding S fixed and restoring faithful forms — no extra features — raises the radius 1.2–6.7× unsigned, which reads as form fidelity being the cheaper axis. **Under arm D it buys** 1.02**–**1.12×: the form price *is* $e_i^{\mathcal{C}}$, the term the signed bound localises. Backbone, forms and features all collapse the same way (App. A43). Clustered by dataset — four clusters, the unit §4 argues for — the bound is worth a geometric-mean $\mathbf{7.9}\times$ $[2.3, 26.7]$ against the knob's $\mathbf{1.48}\times$ $[1.37, 1.61]$: wide against narrow, not overlapping, and it **replicates on four further datasets** (d=8–57): the bound

is worth 6.2–367× against 1.28–1.55×, widening with d — arm A collapsing, not arm D growing (App. A26). At n=8 we still read the ordering, not the magnitude. That prices the knob at one step, which flatters us: over its whole plotted range it is 3.3× (a floor — 49–70% of those instances sit at $r_{\max}$), so the honest factor is about two. **The two are substitutes and Fig. 1 carries the crossing**: on `heart` the bound is worth 19.1× at the minimal explanation and 1.14× by ε_{sel}=0.20, where the knob has bought 2.7×. It reads 1.01× at the tightest budget we plot, where the knob has bought 3.8×, but 67% of *both* arms sit at $r_{\max}$ there, so by the rule we apply to our own MIP replicate that last band is a property of the ceiling. *At minimality* the bound dominates; buy enough features and the two meet, because an explanation naming almost every feature has almost no excluded residual left to cancel.

Post-hoc explanations do move on that ball. A certificate is worth having only if what it forbids happens. On 100 instances across four datasets (mean certified radius 0.21–0.31, five-seed means) the shown explanation stays valid on all 4000 perturbations drawn from the certified balls — a soundness check, not a baseline column — while the top-k *set* moves for every explainer of the same VIB-NAM. There we do not win: exact additive Shapley 0.71, KernelSHAP 0.70, a LIME-style ridge surrogate (Ribeiro et al., 2016) 0.56, our own re-selection 0.62, *worse* than exact Shapley on all five seeds (-0.09, $[-0.12, -0.05]$). Exact Shapley samples nothing and still moves (29% flips), so fragility does not need sampling (the 0.8-point gap to KernelSHAP does not show they are equal) — which is why we certify validity rather than chase a ranking (App. A54). **And ε-validity is not decision invariance**, which a deployer should read first: scored against Bassan et al. (2026)'s sufficiency predicate on our own models, the minimal certified explanation preserves the decision on 29–90% of instances (72/29/90/88), the stabilised one on 42–97%; on `wine` most fail (App. A38).

Part of the d-collapse is the threat model. Where the L_∞ certificate has collapsed, the same explanations certify 1.7–2.5× more radius over an L_2 budget, in which excluded coordinates compete for one budget (arm A both sides; App. A45).

On the deployed model the obstacle is the vocabulary, not the bound. With pairwise terms in the model, a one-dimensional surrogate must absorb every h_{ij} as residual. Letting the explanation display all 20 pairs exactly restores $r \geq 0.1$ to 91–98%, the main-effects level; but on `wine` that takes twelve pairs, when the explanation already names 10.2 of 11 features — accuracy from interactions and a compact certified explanation are in tension there, and the certificate measures that trade rather than removing it (App. A24).

Diagnostics and interpretability checks (summary). On synthetic data (d=10, 10 seeds; the feature-type mix is in App. A3) the IB separates additive-shape from curved features in **10/10 seeds**, as Cor. 1 predicts (Fig. 6); the strict noise->-linear ordering, outside its scope, holds 3/10. At matched fidelity VIB-X names about half the features a full attribution would (App. A32) and beats a same-model LIME baseline on deletion-test AOPC (Samek et al., 2017) by 8.4% (App. A31).

6 Conclusion

Minimality and stability trade against each other, and the trade is cheap to certify — but the exchange rate is set by the bound, not the budget. Tightening certification from the explainer's unsigned acceptance test to the signed residual it displays is free and worth 2.4–37.8×; buying 0.6–1.1 extra features is worth 1.1–1.3× on top (3.2–19.1× and 1.4–1.6× off the censoring ceiling; App. A49); clustered by dataset the two effects are 7.9× [2.3, 26.7] against 1.48× [1.37, 1.61]. Priced in *features* at a matched radius the gap is small, and we state it rather than let the ratio carry the impression: to certify r=0.3 the unsigned criterion needs 0.1–1.3 more features than the signed one. The unsigned frontier stays, labelled, as the explainer's own acceptance certificate. The backbone spread, off the ceiling, is up to 11× unsigned and 1.3–1.4× signed, and we read the collapse, not the ordering: the bound dominates the backbone too. **In practice**: select with magnitudes, but certify against the signed, localised residual before buying features — free, and at $r \geq 0.1$ it turns 2–53% certified into 93–99% (App. A49). **Limitations**: certifying the interaction-bearing model costs **0.44**× the radius at the output level, 0.30× at the explanation level; displaying the pairs restores it, but on `wine` only with 12 of 20 shown (App. A24); $|S|$ grows with d; Thm 1 holds on 7/45 cells; we certify the displayed surrogate's *validity*, not its identity; we ran no user study; and an L_∞ box ignores categorical and correlated features.

Reproducibility Statement

Architecture, training schedule and hyperparameters are in §3, App. A3 and App. A4; every theorem, lemma, corollary and proposition is proved with its assumptions stated inline in Apps. A19–A29 and A49; the protocol defects we disclose, and which run each number comes from, are in App. A9 and App. A52. The anonymized supplement ships the code, the exact per-fold nested-CV split indices, and every result artifact that backs a reported number, with a reproduction command per artifact in `REPRODUCE.md`. A provenance gate (`scripts/ci_provenance_check.sh`) recomputes every headline number, thirty-two of the fifty-one tables and five of the twelve figures from those artifacts on each build and refuses the paper on a disagreement; App. A25 says what it checks, what it cannot, and where it has failed us.

Ethics Statement

The UCI/OpenML benchmarks are public and non-sensitive. The MIMIC-IV (Johnson et al., 2023) / MIMIC-CXR (Johnson et al., 2019) analysis — run, and no longer printed in this revision, though its artifacts ship — used PhysioNet credentialed data under its Data Use Agreement; we treated it as a *retrospective, subject-level* representability stress test, make *no* clinical-deployment or temporal-prediction claim, and the released supplement contains *no* restricted records, subject identifiers, or derived embeddings — only aggregate per-seed scores. That is enforced rather than asserted: `scripts/check_restricted_data.py` runs in the provenance gate and fails the build if any array container or any record identifier (`subject_id`, `hadm_id`, `stay_id`, ...) is tracked for release. Reproducing the MIMIC arms requires the reader's own PhysioNet credentials and the embedding scripts we ship. No human-subjects experiment was conducted.

AI Use Statement

Generative AI (Claude) was used substantially in the *verification and revision* phase of this work, and we disclose that against the categories the ICLR 2027 AI Policy marks as requiring disclosure. The original research — the architecture, the idea of certifying an additive model's reachable interval exactly, the VIB-X formulation, and the first round of experiments — was designed and carried out by the authors without AI assistance.

Uses requiring disclosure. In the revision phase AI was used to *design and give feedback on experiments*: the five-seed replications, the depth$\times$parameter cell, the MIP-optimality study of App. A41, the d-scaling sweep of App. A40, and the comparison against the abductive objective of Bassan et al. (2026) were AI-proposed and AI-implemented. It was used to *implement methods* (the certifier's coverage and padding corrections, the stability-radius bisection, the provenance gates, and the experiment runners this paper cites), to *refine hypotheses* (the framing of minimality as the *cause* of explanation fragility, and the stability–simplicity frontier as its remedy, were sharpened in AI-assisted analysis of the authors' earlier observation that minimal explanations certified near-zero radii), and to *interpret results*. Proof assistance was limited to restating and tightening arguments the authors had established; no theorem in this paper originates with AI. No AI-generated synthetic data underlies a reported result except the controlled sweeps described in App. A40, and no dataset cleaning or translation was involved.

Uses where disclosure is recommended. AI drafted and edited substantial parts of this paper, including its present title and section structure, produced figures and code, and identified and formatted literature.

Review and responsibility. All AI-assisted work was reviewed by the authors, and the review was mechanised where it could be. Every headline number the gate registers is regenerated from a shipped artifact and re-checked against the text on every build (`scripts/check_paper_numbers.py`); `--list` prints that registry, which is 831 numbers and six cross-site claims, not every figure in the paper. Every external citation was verified against a primary source (DOI/arXiv/official proceedings): author list and order, title, venue and year against the publisher's page, the published venue preferred over the preprint wherever one exists, and the version recorded for the entries that exist only on arXiv — the build refuses an arXiv-only entry that pins no version, and refuses one entered as a preprint while naming a venue. Claims that did not survive checking were withdrawn rather than softened: a β–conditional-MI bound, an apportionment of the depth effect, and the applicability of Theorem 1 to our fitted models were all retracted on measurement. **The authors take responsibility for the final**

content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4699–4711. Curran Associates, Inc., 2021. URL https://arxiv.org/abs/2004.13912.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2018. URL https://arxiv.org/abs/1806.08049. arXiv:1806.08049v1; non-archival workshop.

Maksym Andriushchenko and Matthias Hein. Provably robust boosted decision stumps and trees against adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.03526.

Shahaf Bassan and Guy Katz. Towards formal XAI: Formally approximate minimal explanations of neural networks. In *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, volume 13993 of *Lecture Notes in Computer Science*, pp. 187–207, 2023. doi: 10.1007/978-3-031-30823-9_10.

Shahaf Bassan, Michal Moshkovitz, and Guy Katz. Additive models explained: A computational complexity approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2510.21292.

Shahaf Bassan, Yizhak Yisrael Elboher, Tobias Ladner, Volkan Şahin, Jan Kretinsky, Matthias Althoff, and Guy Katz. Provably explaining neural additive models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2602.17530.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Kouroche Bouchiat, Alexander Immer, Hugo Yèche, Gunnar Rätsch, and Vincent Fortuin. Improving neural additive models with Bayesian principles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 4416–4443, 2024. URL https://arxiv.org/abs/2305.16905.

Chun-Hao Chang, Rich Caruana, and Anna Goldenberg. NODE-GAM: Neural generalized additive model for interpretable deep learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.01613.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Alessandro De Palma, Greta Dolcetti, and Caterina Urban. Faster verified explanations for neural networks. In *40th European Conference on Object-Oriented Programming (ECOOP 2026)*, volume 372 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 3:1–3:32, 2026. doi: 10.4230/LIPIcs.ECOOP.2026.3.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490, 2021. URL https://arxiv.org/abs/2108.04884.

Martin E. Dyer. An O(n) algorithm for the multiple-choice knapsack linear program. *Mathematical Programming*, 29(1):57–63, 1984. doi: 10.1007/BF02591729.

James Enouen and Yan Liu. Sparse interaction additive networks via feature interaction detection and sparse selection. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2209.09326.

Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008. doi: 10.1214/07-AOAS148.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. In *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018. arXiv:1807.01622v1; non-archival workshop.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019. doi: 10.1609/aaai.v33i01.33013681.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations (ICLR)*, 2025. Preprint arXiv:2410.24210v3.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. TabTransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020. URL https://arxiv.org/abs/2012.06678. arXiv:2012.06678v1; no peer-reviewed venue.

Xuanxiang Huang and Joao Marques-Silva. From robustness to explainability and back again. *arXiv preprint arXiv:2306.03048*, 2023. URL https://arxiv.org/abs/2306.03048v3. arXiv:2306.03048v3 (2024-12-03), the revised version.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1511–1519, 2019. doi: 10.1609/aaai.v33i01.33011511.

Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *International Journal of Approximate Reasoning*, 159:108939, 2023. doi: 10.1016/j.ijar.2023.108939. URL https://doi.org/10.1016/j.ijar.2023.108939.

Yacine Izza, Xuanxiang Huang, Antonio Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva. Distance-restricted explanations: Theoretical underpinnings and efficient implementation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pp. 475–486, 2024a. doi: 10.24963/kr.2024/45.

Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and Joao Marques-Silva. Delivering inflated explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12744–12753, 2024b. doi: 10.1609/aaai.v38i11.29170. URL https://ojs.aaai.org/index.php/AAAI/article/view/29170.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. URL https://arxiv.org/abs/1901.07042v5. arXiv:1901.07042v5; the JPG release has no peer-reviewed venue (the companion MIMIC-CXR paper is Sci. Data 6:317).

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.05761.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004. doi: 10.1103/PhysRevE.69.066138.

Emanuele La Malfa, Rhiannon Michelmore, Agnieszka M. Zbrzezny, Nicola Paoletti, and Marta Kwiatkowska. On guaranteed optimal robust explanations for NLP models. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2658–2665, 2021. doi: 10.24963/ijcai.2021/366.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-Arnold networks. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2404.19756v5.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 623–631, 2013. doi: 10.1145/2487575.2487579.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.07874.

Jan Macdonald, Stephan Wäldchen, Sascha Hauch, and Gitta Kutyniok. A rate-distortion framework for explaining neural network decisions. *arXiv preprint arXiv:1905.11092*, 2019. URL https://arxiv.org/abs/1905.11092. arXiv:1905.11092v1; no peer-reviewed venue.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability, 2019. arXiv:1909.09223v1; no peer-reviewed venue.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8414–8426, 2022. URL https://arxiv.org/abs/2205.14120.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 1527–1535, 2018. doi: 10.1609/aaai.v32i1.11491.

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.

Sören Schleibaum, Anton Frederik Thielmann, Julian Teusch, Benjamin Säfken, and Jörg P. Müller. Evinam: Intelligibility and uncertainty via evidential neural additive models. *arXiv preprint arXiv:2601.08556*, 2026. URL https://arxiv.org/abs/2601.08556. arXiv:2601.08556v1; no peer-reviewed venue.

Prabhakant Sinha and Andris A. Zoltners. The multiple-choice knapsack problem. *Operations Research*, 27(3):503–515, 1979. doi: 10.1287/opre.27.3.503.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 180–186, 2020. doi: 10.1145/3375627.3375830.

I. M. Sobol. Sensitivity analysis for non-linear mathematical models. *Mathematical Modelling and Computational Experiment*, 1(4):407–414, 1993. English translation of Sobol' (1990), *Sensitivity estimates for nonlinear mathematical models*, Matematicheskoe Modelirovanie 2(1):112–118. Secondary indexes disagree about which of the two titles belongs to this 1993 record; we cite the translation by its venue, volume, issue and pages, which are not in dispute.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021. URL https://arxiv.org/abs/2106.01342. arXiv:2106.01342v1; no peer-reviewed venue.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999. Preprint: arXiv:physics/0004057.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1705.04977.

Stephan Wäldchen, Jan Macdonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70: 351–387, 2021. doi: 10.1613/jair.1.12359. URL https://jair.org/index.php/jair/article/view/12359.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29909–29921, 2021. URL https://arxiv.org/abs/2103.06624.

Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition, 2017. ISBN 978-1-4987-2833-1.

Min Wu, Haoze Wu, and Clark Barrett. VeriX: Towards verified explainability of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2212.01051.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2002.12920.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1901.09392.

Eitan Zemel. An O(n) algorithm for the linear multiple choice knapsack problem and related problems. *Information Processing Letters*, 18(3):123–128, 1984. doi: 10.1016/0020-0190(84)90014-0.

Wei Zhang, Brian Barr, and John Paisley. Gaussian process neural additive models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 16865–16872. AAAI Press, 2024a. doi: 10.1609/aaai.v38i15.29628. URL https://ojs.aaai.org/index.php/AAAI/article/view/29628.

Xinyu Zhang, Julien Martinelli, and S. T. John. Challenges in interpretability of additive models. In *XAI Workshop at the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024b. URL https://arxiv.org/abs/2504.10169.

A1 Appendix Roadmap: What Supports a Claim, and What Does Not

This appendix is long, and not all of it is evidence for the paper's claims. Saying which is which is cheaper for a reader than making them infer it.

The final claims, in one place. Every claim the paper makes, the appendix that establishes it, and its status in this revision. A reader who wants only the result can stop after this table; the rest of the appendix is the evidence and, where marked, the record of what was withdrawn.

Table 1: Claims ledger. "Holds" = supported by the named appendix, under the protocol that appendix states.

Claim	Where	Status
Sound per-input L_∞ radius for continuous additive shapes	A21	holds (exact arithmetic)
Arm D is the attained worst case for main-effects models	A49	holds; sound, not attained, with interactions
At minimality the bound is worth far more than the budget	A49, A48	holds; survives centring (A50) and a fixed-explanation cap
Beyond minimality the two levers meet	§5.3	holds (`heart`: crossing by ε_{sel}=0.20)
The finding transfers to four further datasets	A26	holds (arm A floors disclosed)
Fragility belongs to the objective, not the search	A41	holds on average on four datasets, incl. best-radius tie-breaking; single ties differ up to 328×
Radius vs. a CROWN-verified MLP	A39	holds as selected; at matched accuracy where available
Displaying pairwise terms restores the deployed model's certificate	A24	holds with all 20 pairs shown; `wine` needs 12 (pre-set criterion failed)
Our fidelity predicate is stronger than regression sufficiency	A38	holds: the empty explanation meets theirs at $r \geq 0.1$ on 87–100%
Explanation *identity* is certifiable	A53	only on a far smaller ball
VIB-X near-optimal with a per-instance LP certificate	A36	holds, for the conservative surrogate
Backbone ordering under arm D	A46	withdrawn (not stable)
The bottleneck adds accuracy or radius	A55	withdrawn (fails preregistered test)
Graceful degradation describes our fitted models	A19	withdrawn (hypothesis holds on $7/45$ cells)
β–MI bound; representability lower bound $J/(8L^4)$	A29, A28	withdrawn (false; counterexamples given)
Certified explanations help people understand or audit a model	—	not evaluated (no user study)
An L_∞ box is the right perturbation model for a deployment	—	not evaluated; categorical and correlated features change the admissible set

Load-bearing. These support a statement in the body and a reader checking the paper should start here. **Start with App. A49**: it reports the bound the title is named after and supersedes the criterion every other appendix is measured under. Then A53 (can we certify the explanation is the *same*, not merely valid), A48 (the frontier with both disclosed protocol defects repaired at once, which is what the body quotes) and A51 (`bike` is an hourly series and we split it randomly). Apps. A21 and A19 (the certificates' proofs); A54 and A43 (the frontier, and the second axis we had not priced); A40, A45 and A44 (how it scales, which half of the collapse is the threat model, and what the in-sample validation split cost); A41 (the exact optimum of the same cost is no less fragile, so the fragility is the objective and not our search); A42 and A46 (the mechanism is universal, and the headroom that looked backbone-specific is the bound's); A38 (position against the abductive line); A39 and A37 (the certified-radius and accuracy numbers); A4, A3 and A9 (what was run, and the protocol defects we disclose); A23 (the 2-D block certificate, and how loose the summed model interval is), A26 (the headline on four datasets the frontier was not developed on) and A25 (what the provenance gate checks, what it cannot, and where it has failed us).

Recorded, but not load-bearing. These document work the paper deliberately does *not* claim, kept because the negative and the null results in them are part of the record. Differentiability lets an additive head train *through* a deep encoder, which EBM cannot do. **We ran that line and do not claim it, and this revision no longer prints it.** The advantage belongs to differentiable *additivity* rather than to our bottleneck — a plain-NAM and a NODE-GAM head match it — and we beat a frozen-encoder GA2M in $\mathbf{0/12}$ cells and its mean is higher in 11 — a count whose table this revision cut, so we state it as a claim we no longer show the evidence for; the artifacts remain in `results/multimodal_real/`. A MIMIC case study was a representability stress test on credentialed data, not a clinical benchmark. Six appendices of protocol for that thread, and the overview figure built on it, are cut here for length; the runners and every artifact remain in the repository, and App. A27 keeps the coupling diagnostic, which estimates the gap Thm. 4 identifies. App. A57 and App. A29 record a withdrawn bound and what replaced it. None of the certificates depends on any of this.

A2 Main-Benchmark Figures (relocated to fit body)

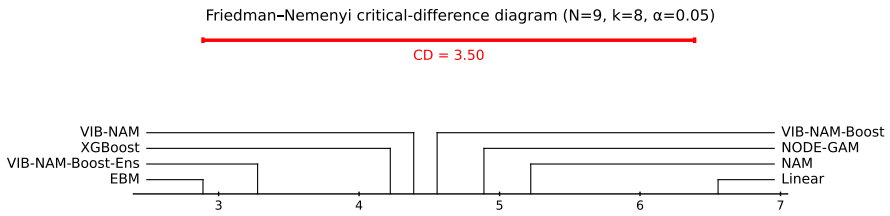

Figure 2: Critical-difference diagram under the nested-CV protocol (8 methods, N=9, $\alpha = 0.05$, CD≈3.50). EBM, VIB-NAM-Boost-Ens, VIB-NAM, NODE-GAM, and XGBoost cluster within CD; only Linear is clearly separated.

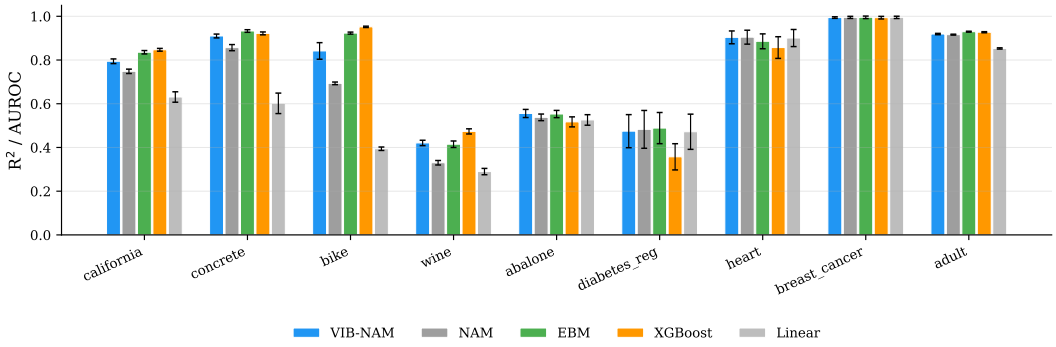

Figure 3: Performance comparison across nine datasets under the nested-CV protocol (mean R^2 / AUROC; outer fold held out for reporting). VIB-NAM is the strongest non-tree baseline and joins EBM / VIB-NAM-Boost-Ens / NODE-GAM / XGBoost in the top-five cluster within Nemenyi's critical distance. Higher is better on all axes.

A3 The Testbed: Architecture, Training, and the Bottleneck's Scope

Relocated from the body: none of this is required to read §4–§5.3, whose constructions apply to any additive model. It is here so that the testbed our numbers were measured on is fully specified.

ARCHITECTURE AND OBJECTIVE

VIB-NAM decomposes predictions as $\hat{y} = b + \sum_i f_i(x_i) + \sum_{(i,j)\in\mathcal{E}} h_{ij}(x_i, x_j)$ with

$$f_i(x_i) = g_i^{\mathrm{nn}}(x_i; \beta_i) + w_i x_i + b_i, \tag{3}$$

where $w_i x_i + b_i$ is an unconstrained linear bypass and g_i^{nn} passes x_i through an encoder (MLP $(256, 256, 128)$, GELU), a stochastic bottleneck $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2 I)$ with d_z=32, and a residual decoder. The loss adds per-feature IB penalties to the task loss, $\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \sum_i \beta_i D_{\mathrm{KL}}[q_\phi(z_i|x_i) \,\|\, p(z_i)]$ with $p(z_i) = \mathcal{N}(0, I)$, so the KL takes the usual Gaussian closed form. Large β_i compresses toward the prior; small β_i preserves nonlinear capacity.

Why β is swept, not learned. Treating β_i as a free parameter is degenerate: since $D_{\mathrm{KL}}[\cdot \,\|\, \cdot] \geq 0$ we have $\partial\mathcal{L}/\partial\beta_i = D_{\mathrm{KL}}[\cdot \,\|\, \cdot] \geq 0$, so gradient descent drives every β_i to its floor with no counteracting term, and a freely optimised β carries no data-driven meaning. We therefore read the resulting channel rate rather than treat β as an estimate, and we fix β on a grid wherever a specific value is the object of study — the certified-radius sweep of §4 (`run_exact_certification.py`, `freeze_beta`). It is *not* fixed in the default four-phase schedule that produces Table 24, the frontier and the signed-criterion runs: there $\tilde{\beta}_i$ trains, as phase (2) below says. An earlier revision wrote "swept, never learned … fixed throughout", which contradicted both the schedule on this page and the runners; an audit caught it. A genuinely *meaningful* learned per-feature rate needs a constrained rate budget or a target-rate Lagrangian, which we leave to future work.

FOUR-PHASE TRAINING AND INTERACTION DISCOVERY

Premature β updates collapse bottlenecks before shape functions learn. We use a four-phase schedule (default epochs): **(1) Warmup** (60): β_i fixed at 0.1, shape functions trained for accuracy; **(2) Anneal** (120): KL weight ramps $0 \to 1$ while $\tilde{\beta}_i$ trains; **(3) Interactions** (150): pairwise h_{ij} fit residuals $r = y - \sum_i f_i(x_i)$ with main effects frozen; **(4) Joint fine-tune** (60): all parameters optimized together. **Interaction discovery.** After Phase 2 we screen pairwise interactions on the residual $r = y - \sum_i f_i(x_i)$ by Pearson correlation $|\mathrm{corr}(r, x_i x_j)|$ and retain the top k=20 pairs (an optional $\rho_{\min}$ threshold and a stricter KSG-MI test (Kraskov et al., 2004) with BH-FDR (Benjamini & Hochberg, 1995) are available). Retained pairs are modeled by a 2D lookup table h_{ij} and trained in Phase 3.

THEOREM: GRACEFUL DEGRADATION

A key property of the linear bypass is that it prevents information collapse under high compression.

Theorem 1 (Graceful Degradation, conditional form; informal)**.** *Let $f_i = g_i^{\mathrm{nn}}(\cdot; \beta_i) + w_i x_i + b_i$ with* Dec *continuous in the latent code, and assume that at each β_i the fitted parameters achieve total objective $\mathcal{L}_{\mathrm{pred}} + \beta_i D_{\mathrm{KL}}^{(i)}$ no worse than the bypass-only reference, and that the decoders are L_D-Lipschitz uniformly in β_i. Then the posterior-mean nonlinear path is within $L_D\sqrt{2C/\beta_i}$ of a constant in $L^2(P_i)$, so f_i is within that distance of an affine function: the shape function degrades from nonlinear to* linear*, not to a constant. Formal statement and proof: App. A19.*

The mechanism is the Gaussian posterior collapsing to the prior, so $g_i^{\mathrm{nn}} \to c_i$ and the bypass survives — unlike a standard NAM, whose f_i collapses to a constant. **The hypothesis is about optimisation, and it mostly fails**: over all 45 (dataset, β) cells it holds on $\mathbf{7/45}$ — $7/9$ datasets at β=0, **none** at any $\beta > 0$, because early stopping selects on prediction loss alone. The theorem therefore says nothing about our fitted models at the β we sweep; synthetic recovery (§5.3, $\bar{\beta}$=0.034/0.035/0.020) is evidence for its *conclusion*, never its hypothesis (App. A19).

Corollary 1 (Channel rate is a nonlinearity test; informal)**.** *Under a local quadratic-risk assumption, $R_i^\star$=0 $\iff$ $f_i^\star$ is affine in the small-β regime, because the bypass absorbs the affine part exactly. Statement and proof: App. A20.*

It recasts §5.3: noise *and* linear features both sit in the $R^\star$=0 class, so only curved ones keep positive rate. The $10/10$ curved-vs-additive separation is its exact prediction; the $3/10$ noise-vs-linear ordering is outside its scope — and we flag that the out-of-scope half is the half that fails.

Scope of the Bottleneck

Two further results sit in the appendix because neither carries a claim of this paper: Thm 4 (representability gap, App. A28), which underwrites the multimodal observation we explicitly do not claim; and Prop. 3, which replaces the β–conditional-MI bound one might hope for and which **is false** — the linear bypass makes $X_i \to Z_i \to f_i(X_i)$ a non-Markov chain, so $I(X_i; f_i(X_i))$ is maximal regardless of β_i. What survives is that the bottleneck constrains the *nonlinear residual channel only* (App. A29), and every β statement here has that scope; Zhang et al. (2024b) urge the same restraint generally, on nonidentifiability grounds we do not address.

A4 Hyperparameter Settings

Which experiments use which configuration. Three configurations exist, and Table 2 is only one of them, so we name all three. (i) The *canonical v2 runner* (`run_full_benchmark.py`) produces the accuracy benchmark and is what the table lists. (ii) The *certification* study (`run_exact_certification.py`) uses the same $(256, 256, 128)$ capacity but with `auto_interactions=False`: when these runs were made the 2-D path was unpadded and the certifier refused a sound certificate over it. Prop. 1 has since removed that restriction (App. A23 certifies the interaction-bearing model directly); this family is reported as it was run. (iii) The *explanation* studies train $(128, 128)$ encoders with d_z=16 for 60 epochs — all seven of them, namely `run_{explanation_stability, stability_frontier, form_fidelity, shapley_topk_control, scale_frontier, mip_fragility, abductive_comparison}.py`. The frontier is therefore a statement about a smaller model than the benchmark reports. All three use the same G=10^6 grid knots; at 10^5 the padding is tenfold larger and the reported fragility inflates with it (App. A54).

The in-sample validation split, and what it costs. All *eight* explanation runners — the seven above plus `run_backbone_headroom.py`, the producer of App. A46 — pass the first 200 training rows as the validation set, so early stopping there is in-sample. On `heart` it is worse than in-sample: the training split is 189 rows, so `Xtr[:200]` returns *the entire training set* and the validation set and the training set are the same object (our artifacts record `n_train: 189`; Table 35's $151 + 38$ says the same). `diabetes_reg` is 65% in-sample by the same arithmetic. `run_stability_frontier.py --holdout_val 0.2` rebuilds the entire frontier on a genuine $80/20$ train/validation partition of the training rows — five seeds, same G=10^6 grid, same evaluation points. The frontier survives: budget consumption at ε_{sel}=ε moves from 85.2–98.7% to 87.2–98.8%, explanation sizes move by at most 0.43 features, and the radius still rises monotonically as ε_{sel} tightens on every dataset in both arms. App. A44 gives the full comparison, including the one cell that moves materially.

Table 2 lists the hyperparameters of the canonical v2 runner (`elasticnam/models/elastic_nam_v2.py` and `elasticnam/experiments/run_full_benchmark.py`). The table is complete and the initialisation is stated as the code computes it, $\mathrm{softplus}(-2)$.

Baselines: EBM runs at library defaults, which means `interactions="5x"` — a GA2M (Lou et al., 2013), not a main-effects GAM. NODE-GAM does *not* run at library defaults; its deviations are listed in App. A9. XGBoost uses 300 estimators, depth 6, learning rate 0.1, subsample 0.8, L2 regularization 1.0. NAM (*matched-capacity* variant) uses the identical $(256, 256, 128)$ encoder but with no IB, no bypass, no interactions; standalone NAM uses $(128, 64, 32)$ per-feature MLPs.

A5 Dataset Details

All datasets are from the UCI ML Repository or scikit-learn. Features are standardized using `StandardScaler` (zero mean, unit variance) for all tasks. Standardized values are clipped to $[-6, 6]$ to reduce the influence of outliers. Table 3 summarizes the datasets.

One of them is a time series and we do not split it as one. `bike` (OpenML 42712) is the UCI Bike Sharing *hourly* file — one row per hour over two years, with `year`, `month`, `hour` and `weekday` among its features and a lag-1 target autocorrelation of $+0.844$. Every split in this paper shuffles, so

Table 2: Hyperparameter settings for VIB-NAM (canonical v2 runner).

Parameter	Value	Description
Encoder hidden dims	$(256, 256, 128)$	MLP hidden layers, GELU
Bottleneck dim d_z	32	Latent dimension
Decoder layers	$(128, 64)$ + residual	Decoder, GELU
Learning rate (weights)	1×10^{-3}	Adam optimizer
Learning rate ($\tilde{\beta}_i$)	5×10^{-4}	separate LR for IB params (v2)
β_{init}	$\text{softplus}(-2)$=0.127	Initial IB value
Dropout	0.1	Encoder and decoder blocks
Warmup epochs T_1	60	Phase 1
Anneal epochs T_2	120	Phase 2
Interaction epochs T_3	150	Phase 3 (main effects frozen)
Fine-tune epochs T_4	60	Phase 4, at $0.05\times$ the base LR
Max epochs / patience	300 / 30	Early stopping, phases 1–2
Gradient clip / weight decay	0.5 / 1×10^{-5}	All phases
Batch size	1024	All phases
Interaction prescreen top-k	20	Pearson top-k pairs ($\rho_{\min}$ optional)
Interaction grid n_{bins}	32–64	2D lookup resolution (64 canonical, 32 for boost)
VIB-X tolerance ε	0.1–0.5	Analysis / default explanation

adjacent hours land on both sides of every fold. App. A51 measures what that costs: every arm loses 0.08–0.15 R^2 under a contiguous temporal split, `bike`'s own ordering changes on 3 of 8 arms, and the average ranks over the nine datasets move by at most 0.33.

Table 3: Dataset statistics.

Dataset	N	d	Task	Source
california	20,640	8	Regression	scikit-learn
heart	270	13	Classification	UCI
breast_cancer	569	30	Classification	scikit-learn
diabetes_reg	442	10	Regression	scikit-learn
concrete	1,030	8	Regression	UCI
bike	17,379	12	Regression	UCI
wine	6,497	11	Regression	UCI
abalone	4,177	8	Regression	UCI
adult	48,842	14	Classification	UCI

A6 Extended Classification Benchmark

To probe generalization beyond the 9 main datasets, we evaluate on three additional mid-sized OpenML classification datasets (bank, magic, credit_default). The reported numbers in Table 4 were collected under the legacy validation-as-test protocol; VIB-NAM trails EBM by 1–2% AUROC on bank/magic and is within 0.002 on credit_default, consistent with the top-cluster picture in the main benchmark. A nested-CV rerun for the extended set is left to a nested-CV follow-up.

Table 4: Extended classification benchmark (5-fold AUROC).

Dataset (N)	VIB-NAM	VIB-NAM-B	VIB-NAM-B-Ens	NAM	EBM	XGBoost	Linear
bank (45k)	0.922	0.922	0.922	0.916	**0.933**	<u>0.934</u>	0.873
magic (19k)	0.917	0.916	0.916	0.906	<u>0.937</u>	**0.938**	0.839
credit_default (30k)	0.782	0.782	0.782	0.779	**0.784**	0.772	0.724

A7 DEEP-TABULAR BASELINE COMPARISON

Table 5 reports nested-CV performance (same outer 5-fold KFold, inner 20% validation split as the main benchmark) for three deep-tabular methods. All three use the same data pipeline as VIB-NAM (StandardScaler, clip $[-5, 5]$, seed 42). TabM uses k=32 ensemble heads, 3-block MLP backbone, lr=10^{-3}. SAINT uses 1 col+row attention block pair, embedding dim 64, batch 64, lr=3×10^{-4}. TabTransformer uses 3-layer Transformer encoder, embedding dim 32, batch 128, lr=10^{-4}. TabTransformer shows instability on diabetes_reg (N=442), where the transformer is severely over-parameterised relative to dataset size; this is a known limitation of deep tabular models on small-N regression. Scripts: `run_tabm_baseline.py`, `run_saint_baseline.py`, `run_tabtransformer_baseline.py`. Results: `results/tabm_baseline/tabm_seed42.json`, `results/saint_baseline/saint_seed42.json`, `results/tabtransformer_baseline/tabtransformer_seed42.json`.

Table 5: Deep-tabular baselines vs. VIB-NAM under the same nested-CV protocol (R^2 for regression, AUROC for classification; seed 42). VIB-NAM's column is `results/benchmark/full_benchmark_seed42_nested.json`, the same artifact as Table 24. **Not a matched-compute comparison, in four ways**: the three baselines subsample to 10,000 rows (`max_n`), so on `adult`, `california` and `bike` they see 20–58% of the data VIB-NAM sees; TabTransformer ran at depth 3 against its own script's default of 6 (SAINT's depth 1 *is* its script's default); both ran 100 epochs with patience 12 against VIB-NAM's 300 and 30; and their inputs are clipped to $[-5, 5]$ where VIB-NAM's are clipped to $[-6, 6]$. None received a tuning budget — nor did VIB-NAM (App. A9). Two further handicaps went undisclosed and are worse; see the text after the table. **Bold** = best of the four methods shown. TabTransformer† fails on diabetes_reg (N=442) due to over-parameterisation.

Dataset (task)	VIB-NAM	TabM	SAINT	TabTransformer
california (reg)	**0.794**	0.790	0.782	0.778
concrete (reg)	**0.910**	0.816	0.885	0.832
bike (reg)	0.841	0.940	**0.941**	0.919
wine (reg)	**0.421**	0.389	0.375	0.367
abalone (reg)	0.555	**0.573**	0.565	0.563
diabetes_reg (reg)	**0.475**	0.390	0.451	$-0.716^\dagger$
heart (cls)	**0.904**	0.895	0.899	0.900
breast_cancer (cls)	0.994	**0.996**	0.992	0.995
adult (cls)	**0.919**	0.902	0.901	0.904

Two further handicaps were not disclosed and are more serious than the four that were. First, the `max_n` subsample is drawn *before* the KFold split, so on `adult`, `california` and `bike` these three methods are scored on a different, smaller test set than the VIB-NAM column of Table 24: on 3 of 9 rows the columns are not computed on the same evaluation data. Second, none of the three standardises the regression target — there is no `y_mean`/`y_std` anywhere in their runners — while VIB-NAM and the plain NAM both do. That is the likelier explanation for TabTransformer's -0.716 on `diabetes_reg` (targets in $[25, 346]$, so the squared error starts near 10^4) than the over-parameterisation the caption attributes it to.

We therefore do not offer this table as a comparison. Four disclosed handicaps, two undisclosed ones, no tuning on either side, and three rows scored on non-identical test sets support no inference in either direction; note also that two of the three subsampled datasets are VIB-NAM wins. We keep it because deleting a table that goes partly against us after running it would be worse, and because accuracy is not a claim of this paper. A reader who wants a deep-tabular comparison should run one at parity. On the same grounds the two remaining arms are reported as configuration failures on our side rather than as properties of those methods: TabPFN returned non-finite scores on every fold (63 nulled) and FT-Transformer diverged on three regression datasets (R^2 to -3.65), and TabPFN in particular is strong in exactly this small-N regime.

A8 Extended Ablation Variants

Provenance of Δ_1, and why it changed. The matched-capacity numbers quoted in §5.1 come from `results/matched_baselines/matched_baselines_seed42.json`, produced by `elasticnam/experiments/run_matched_baselines.py`. Three arms at the canonical $(256, 256, 128)$ capacity — a plain NAM, the same network with the VIB path and the linear bypass added, and that with Phase-3 interactions on — so that Δ_1 and Δ_2 are each a *single*-component difference. Scoring is the headline nested CV: outer 5-fold, an inner 20% split of each outer training fold for early stopping only, the outer fold scored once.

Under the legacy validation-as-test scoring this appendix flags as leaky, the same comparison reads $\Delta_1 = -0.003$ and $8/9$; on the nested protocol it is -0.005 and $7/9$, with `concrete` at -0.037 — slightly less favourable to us. The nested-CV ablation below reaches the same conclusion independently, with *no_ib* inside ± 0.009 on all five of its datasets.

Table 6 reports delta in performance relative to the full VIB-NAM for six design variants on five datasets (3-fold CV). Positive values = better than full model; negative = worse. Removing interactions is the largest single effect. The no_ib variant has a mean positive delta ($+0.004$ on regression), which is the honest cost of the IB's *diagnostic* benefit (see Section 5.3). No_warmup hurts concrete (-0.010). Bottleneck size d_z is robust. **Confound disclosure.** The `no_interaction` variant runs Phase 1 only, whereas `full` additionally runs the Phase-3 interaction fit and Phase-4 joint fine-tune; the artifact records this per variant in a `phases_run` field (`results/ablation/ablation_seed42.json`). Its delta therefore conflates removing interactions with removing the later-phase training budget. It is an upper bound on the interaction contribution only if the extra phases never hurt on their own, which we did not test, so we read it as a confounded estimate, not an isolated one and not a bound. Per-fold scores (`per_fold_score`) are persisted for paired re-analysis.

Table 6: Ablation study: performance delta from full VIB-NAM (positive = better than full model; negative = worse).

Variant	california	heart	diabetes	concrete	breast_cancer
no_interaction	-0.050	-0.009	-0.004	-0.051	$+0.000$
no_ib	$+0.009$	-0.000	-0.003	$+0.008$	$+0.001$
no_bypass	$+0.000$	-0.003	-0.002	$+0.001$	-0.002
no_warmup	-0.005	-0.004	-0.008	-0.010	$+0.001$
cosine_anneal	-0.001	-0.002	-0.008	$+0.001$	-0.002
small_bottleneck	$+0.007$	-0.008	-0.003	$+0.001$	-0.000
large_bottleneck	$+0.004$	$+0.005$	-0.002	-0.001	$+0.001$

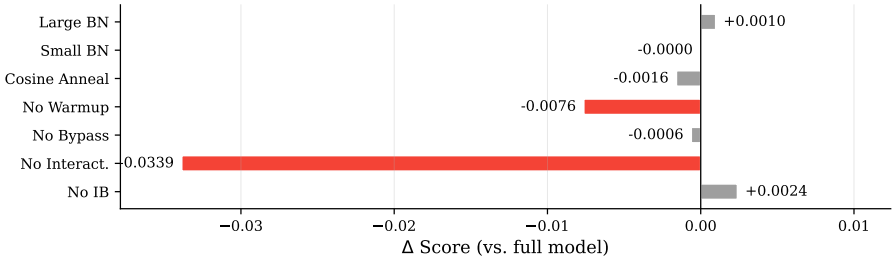

Figure 4: Ablation study: mean Δ score across five datasets relative to the full model.

A9 Evaluation Protocol Disclosure

Table 24 is the *nested* artifact `full_benchmark_seed42_nested.json`, produced by `run_full_benchmark.py --inner_val_split`: outer 5-fold KFold, an inner 20% split of each outer training fold used for early stopping, phase selection and boost-round acceptance, and the held-out outer fold scored once. The single-validation protocol — outer fold used as both validation and test — is superseded and is not the headline benchmark.

What the nested protocol costs us. Recomputed there, VIB-NAM improves over the plain NAM on $\mathbf{6/9}$ datasets, not $8/9$: it loses on `diabetes_reg` (0.475 vs 0.483), `heart` (0.9036 vs 0.9044) and `breast_cancer` (0.9942 vs 0.9950) — margins of under a point in every case, and Table 24 already bolds NAM on `heart`. The $8/9$ figure came from the legacy validation-as-test artifact this appendix says is not the headline. Accuracy is not a claim of this paper (§5.1); it should nonetheless not have been quoted from the artifact we ourselves retired.

The protocol is not identical across arms, and we should not have said it was. EBM, XGBoost, NODE-GAM and Linear are fit on the 64% inner-train rows and never see the inner validation split or get refit on the full outer training fold; NODE-GAM carves a further internal 15% and so sees 54% of N. No method is advantaged on data — the VIB-NAM arms also train on 64% — but four baselines are handicapped relative to a standard nested refit. The model-selection budgets are also asymmetric: the boosted-ensemble arm makes on the order of sixty validation-gated accept/reject decisions per fold and XGBoost makes none. Two baseline descriptions need stating precisely. First, NODE-GAM does *not* run at library defaults, but not where one might look: `num_trees=200`, `num_layers=2`, `depth=3` and `lr=0.01` *are* NodeGAM's defaults. The deviations are all cuts to the training budget: `max_steps=3000` against a default of `20000` for `NodeGAMRegressor` and `10000` for `NodeGAMClassifier`, `early_stopping_steps=300` against `2000`, and `batch_size=1024` against `2048`; the last two were never disclosed anywhere. NODE-GAM therefore gets about $1/6.7$ of its intended budget on the six regression datasets and about a third on the three classification ones, which belongs on the handicap list above and not in a list of architecture choices we made deliberately. Second, EBM at library defaults runs `interactions="5x"`, i.e. a GA2M with $5d$ pairwise terms rather than a main-effects GAM — and `ga2m=1` is likewise a NodeGAM default that we never override, so *both* GAM baselines in Table 24 carry pairwise interactions while the arm we certify in §5.3 is main-effects-only. That is arguably the right opponent for a model with discovered interactions, but it is not what a reader would assume, and it means the accuracy table and the certification study are not reporting on the same function class.

Single seed. The entire accuracy benchmark — Table 24, the ranks, Friedman, Holm, every Wilcoxon, the matched-capacity Δ_1 and the ablation — is seed 42 alone over 5 folds, while the certification and frontier studies use five seeds. Per-fold σ reaches ± 0.0865 (NAM on `diabetes_reg`), and ± 0.0802 for our own headline arm on that same dataset. That dispersion matters: the $8/9 \to 6/9$ correction this appendix reports as its headline cost turns on a `diabetes_reg` gap of 0.008, roughly a tenth of the fold-level σ there. Neither $8/9$ nor $6/9$ is distinguishable from noise, and the corrected figure should be read as a withdrawal of the comparison rather than as a smaller version of it. This paper elsewhere withdraws a claim for being single-seed (App. A39); the same caution applies here and the accuracy table should be read as descriptive.

Other mitigations. Absolute scores may be optimistic relative to a fully held-out protocol. The matched-capacity ablation (§5.1) and the noise-injection IB diagnostic (App A16) are not subject to the comparison-fairness issue because the variants are trained side by side. The VIB-X optimality results (§5.2) measure greedy against exact MIP on the *same* model and are unaffected by the protocol.

Neither arm was tuned. The paper says of the deep-tabular baselines that "none received a tuning budget" (App. A7); the same is true of VIB-NAM. The benchmark's hyperparameters come from `vars(args)`, i.e. the command-line defaults reproduced in Table 2, and the shipped artifact's `cli_args` records exactly those values. `elasticnam/experiments/tune_optuna.py` exists and was used during development, but no benchmark runner reads its output, so no reported number is the product of a search. This removes a leak we had to check for — an Optuna study splits the same X the outer folds later test on — and it also means the accuracy comparison is default-versus-default rather than tuned-versus-default in our favour.

Five-fold Wilcoxon cannot reject, and we do not ask it to. The paired signed-rank statistic is discrete: on n pairs the smallest attainable two-sided p is $2/2^n$, so $n{=}5$ floors at 0.0625 and $n{=}4$ at 0.125 — both above $\alpha{=}0.05$, no matter how large the effect. The per-dataset entries in `results/significance/` are five-fold tests and **35** of them sit at exactly that floor (a thirty-sixth entry at 0.0625 is an *aggregate*, $n{=}9$, whose floor is 0.0039 and whose p merely coincides, so it is a genuine dataset-level test). None of the thirty-five is cited: every inferential claim in this paper is made at the dataset level ($n{=}9$) or at the instance/fold level ($n{=}10$). The artifacts carry

`min_attainable_two_sided_p` and `underpowered_at_alpha_0.05` beside every p, so a reader cannot mistake a floored 0.0625 for a near-miss.

The ceiling that actually binds Table 24. The n=5 floor is the one usually quoted; it is not the one governing our headline table. At the dataset level n=9 gives a floor of $2/2^9 = 0.0039$, and the family is seven baseline comparisons, so the smallest attainable *Holm-adjusted* p is $7 \times 0.0039 = 0.0273$, so only a clean $9/9$ sweep clears α=0.05 at all. Two conclusions that do *not* follow, because both are easy to draw from that arithmetic: significance is not restricted to the top-ranked comparison (under Holm's monotone step-down a *second* $9/9$ sweep is compared against 6×0.0039 and also clears — and `ENAM-Boost` (VIB-NAM's boosted sibling) does sweep $9/9$, carrying `wilcoxon_p_holm` $= 0.0273$ with `significant_after_holm` true); and the two failures we report (0.14, 0.12) were not unreachable by design, since reachability is a function of the observed win count. Those two comparisons won $7/9$, whose smallest attainable p is 0.0195. Only one of them scored it: the signed-rank statistic reads the magnitudes and not just the win count, so `Linear` scored 0.0195 and sits at Holm rank 2 ($\times 6 = 0.117$) while `NODE-GAM` scored 0.0273 at rank 3 ($\times 5 = 0.137$). The correct reading is narrower: the family ceiling leaves 0.0273 of margin below α, so nothing short of a sweep can register, and a $7/9$ result is uninformative rather than negative.

One further caveat on that family. Its seven members include `ElasticNAM` and `ENAM-Boost` — the benchmark artifact's keys for VIB-NAM and its boosted sibling, kept verbatim here because they are what the shipped JSON records — which are our own ablation siblings rather than external baselines. Counting them makes the correction *more* conservative for any claim against a baseline, so it does not flatter us, but it is not the standard construction and a reader comparing our adjusted p against a five-baseline family elsewhere is not comparing like with like.

The nested-CV runner that reserves an inner validation split for selection and reports only on the held-out outer fold is exposed by `run_full_benchmark.py --inner_val_split`; the artifact path suffixed `_nested.json` is the canonical source for Table 24 and the significance tests. The artifact persists `fold_results` (per-fold metric values), `fold_indices`, and a `protocol` field; `run_significance.py` now raises if those per-fold scores are missing, so dataset-mean Wilcoxon cannot silently replace fold-level Wilcoxon. The single-validation artifact (outer fold used as both validation and test) is retained for comparison only and is not the headline.

A10 Full Calibration Table (NP vs. MC-Dropout vs. Deep Ensemble)

Full per-dataset calibration comparison (3-fold CV). We report R^2, Expected Normalized Calibration Error (ENCE, ↓), Prediction Interval Coverage Probability at 95% (PICP_{95}; closer to 0.95 is better), negative log-likelihood (NLL, ↓), and interval sharpness (↓). MC-Dropout uses dropout-based epistemic variance combined with aleatoric variance estimated from *training* residuals (matching the implementation in `run_calibration.py`). Deep Ensemble uses variance across $K = 5$ seeded models (epistemic only).

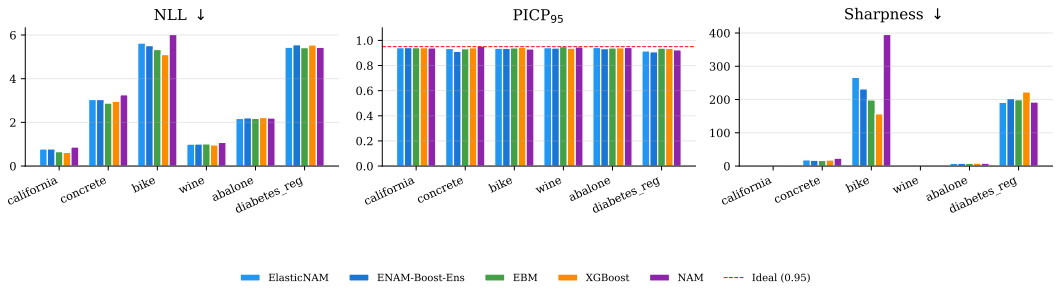

Figure 5: Uncertainty quantification across four regression datasets (3-fold CV). NP and LA-NAM are conservative ($\text{PICP}_{95} \approx 1.0$) but dominated by MC-Dropout with residual aleatoric variance on ENCE/NLL; Deep Ensembles (epistemic only) under-cover. Best per dataset bolded in Table 7.

Table 7: Calibration comparison across four regression datasets (3-fold CV). Best per dataset in **bold**; for PICP_{95} best is closest to the nominal 0.95, so 1.000 is over-coverage, not a win. GP-NAM is an additive per-feature Gaussian-process baseline (cyclic training on residuals, subsample 400; `results/calibration/gpnam_baseline_seed42.json`); LA-NAM row is a Laplace-approximation analogue on top of trained VIB-NAM (App. A17).

Dataset	Method	R^2	ENCE ↓	PICP_{95}	NLL ↓	Sharp. ↓
california	NP	0.716	0.732	1.000	1.773	8.87
	GP-NAM	0.669	0.623	0.999	1.55	6.87
	LA-NAM	0.741	0.837	1.000	2.218	14.61
	MC-Drop	0.739	**0.079**	**0.944**	**0.888**	2.33
	DeepEns	**0.745**	11.30	0.189	96.42	**0.22**
concrete	NP	0.812	0.785	1.000	4.440	129.5
	GP-NAM	0.534	1.239	0.685	5.13	21.70
	LA-NAM	0.850	0.903	1.000	5.124	270.56
	MC-Drop	0.848	**0.104**	**0.965**	**3.291**	25.75
	DeepEns	**0.853**	11.69	0.144	105.4	**2.34**
diabetes	NP	0.480	0.673	1.000	6.088	651.8
	GP-NAM	−0.525	8.339	0.373	96.63	92.50
	LA-NAM	0.487	0.758	1.000	6.414	950.1
	MC-Drop	0.499	**0.166**	**0.932**	**5.418**	202.3
	DeepEns	**0.500**	11.55	0.156	106.0	**23.3**
wine	NP	0.326	0.644	1.000	1.680	7.87
	GP-NAM	0.244	0.661	1.000	1.783	8.78
	LA-NAM	0.339	0.691	1.000	1.802	9.25
	MC-Drop	0.336	**0.071**	**0.945**	**1.079**	2.76
	DeepEns	**0.341**	16.52	0.094	199.1	**0.19**

A11 VIB-NAM-BOOST COMPONENT ABLATION

Table 8: VIB-NAM-Boost-Ens component ablation (ΔR^2 vs. full model). Boosting (*no_boost*) is the dominant contributor; row-subsample bagging hurts small datasets.

Dataset	no_boost	no_cyclic	no_ensemble	bagging	boost_5	boost_15	ensemble_5
california	−0.010	−0.004	−0.002	−0.005	−0.002	−0.001	+0.000
concrete	−0.023	−0.002	−0.000	−0.010	+0.001	−0.001	−0.000
bike	−0.084	−0.001	−0.001	−0.005	−0.002	−0.001	−0.000
wine	−0.022	+0.004	+0.000	−0.025	+0.003	+0.003	+0.002
diabetes_reg	+0.002	+0.010	−0.003	+0.006	−0.010	−0.006	−0.000
Mean Δ	**−0.027**	+0.001	−0.001	−0.008	−0.002	−0.001	+0.000

A12 HYPERPARAMETER SENSITIVITY

A13 DELETION-TEST AOPC

A14 COMPUTE DISCLOSURE

Main-benchmark experiments (5-fold CV, 9 datasets, 7 models) run in approximately 12 GPU-hours on a single NVIDIA RTX-3090 (24 GB). Per-model mean training time per fold (averaged across 9 datasets): VIB-NAM 58s, VIB-NAM-Boost 71s, VIB-NAM-Boost-Ens 214s, NAM 37s, NODE-GAM 185s, EBM 47s (CPU), XGBoost $<$ 1s, Linear $<$ 1s (recomputed from the canonical benchmark artifact; matches §5.1).

VIB-X latency (measured directly per sample, mean over 30 samples per dataset; `results/atlas_latency/atlas_latency_seed42.json`): the greedy heuristic runs in 0.53–

Table 9: Hyperparameter sensitivity (R^2 on three datasets when varying one hyperparameter at a time). Maximum deviation from default ± 0.02.

HP	Values	california	concrete	diabetes
d_z	8	0.805	0.934	0.526
	16	0.807	0.933	0.507
	32 (default)	0.808	0.934	0.508
	64	0.805	0.933	0.507
	128	0.808	0.935	0.514
η	0.1	0.808	0.935	0.510
	0.2	0.805	0.937	0.503
	0.3 (default)	0.807	0.936	0.512
	0.4	0.807	0.935	0.502
	0.5	0.797	0.939	0.503
T	0	0.803	0.933	0.502
	5	0.807	0.934	0.498
	10 (default)	0.796	0.935	0.504
	15	0.808	0.936	0.496
	20	0.807	0.934	0.502
K	1	0.807	0.935	0.506
	3 (default)	0.809	0.935	0.505
	5	0.808	0.935	0.505

Table 10: Deletion-test AOPC scores with 95% bootstrap CIs (B=200 resamples). Higher = more faithful. VIB-NAM attributions outperform random ranking on all four datasets with non-overlapping CIs, which is what the bold marks and all this table supports. The SHAP row is TreeSHAP on a separately trained XGBoost, scored by masking through *that* model: it is ahead here on three of the four datasets and the comparison is not commensurable in either direction (App. A31 says why, and gives the same-model VIB-X/LIME contrast that is).

Method	california	concrete	diabetes_reg	heart
Random	0.623 [0.61,0.64]	0.673 [0.61,0.73]	0.389 [0.35,0.44]	1.803 [1.58,2.01]
VIB-NAM	**0.763 [0.75,0.78]**	**0.821 [0.75,0.89]**	**0.556 [0.48,0.62]**	**2.521 [2.19,2.87]**
SHAP	0.854	0.849	0.566	2.432

0.94 ms/sample across $d \in \{8, \ldots, 13\}$, roughly **11–21**$\times$ faster than the exact PuLP/CBC MIP solver (9.8–14.9 ms/sample) while achieving mean cost ratio ≤ 1.009. Both scale roughly linearly in d in this range. Precise per-dataset numbers: `concrete` (d=8) greedy 0.60 ms / MIP 12.6 ms; `abalone` (d=8) 0.57 / 9.8 ms; `diabetes_reg` (d=10) 0.91 / 14.9 ms; `wine` (d=11) 0.53 / 10.8 ms; `heart` (d=13) 0.94 / 10.7 ms. The grid matters here for soundness rather than speed: $e_i^{\mathcal{C}}$ carries the Lipschitz padding $L_i h/2$, which at G=500 exceeds ε=0.1 for every feature on these models, so constant/linear/quadratic are all infeasible and both solvers would be timed on a degenerate problem with one form per feature (a ratio of 15–30$\times$ results). At the G=200001 the optimality table uses, the MIP has a real choice set and the gap is the one quoted. `ATLASCertifier` raises a warning when the choice set collapses, so the two settings cannot diverge silently.

Parameter counts (analytical). Each VIB-NAM feature sub-network comprises a $(1 \to 256 \to 256 \to 128)$ encoder, twin $(128 \to 32)$ $\mu/\log\sigma^2$ heads, a $(32 \to 128 \to 64 \to 1)$ decoder, and a 2-parameter linear bypass, totalling approximately 120,003 parameters per feature. Full-model parameter counts are therefore: $d = 8$: 0.96M; $d = 10$: 1.20M; $d = 11$: 1.32M; $d = 13$: 1.56M; $d = 30$ (breast_cancer, our largest main-benchmark dataset): 3.60M. For comparison, EBM on the same datasets uses $\sim$ 1M tree-bag parameters per shape, and vanilla NAM with $(128, 64, 32)$ per-feature MLPs totals $\sim$ 12k parameters per feature (96k for $d = 8$). VIB-NAM is therefore $\sim 10\times$ larger in parameters than vanilla NAM and of comparable order to EBM. The matched-capacity ablation (§5.1) controls for this disparity.

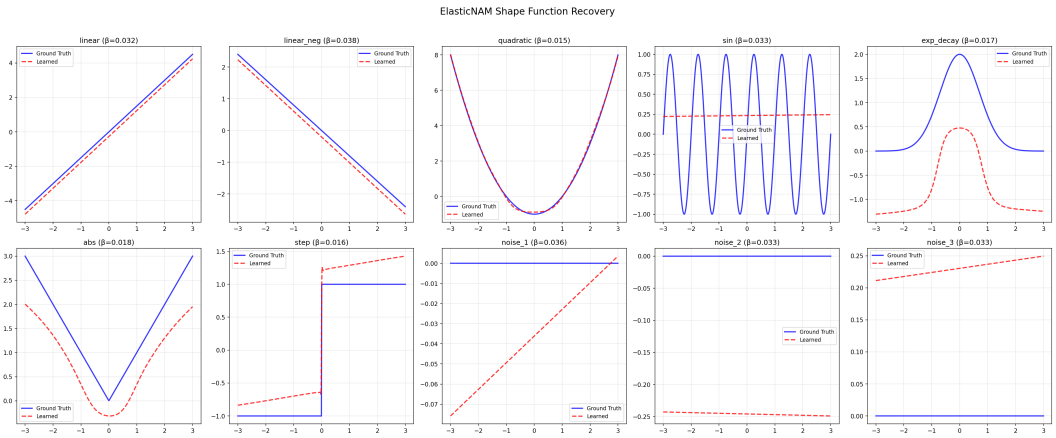

Figure 6: **Learned shape functions on the d=10 synthetic benchmark** (ground truth solid, learned dashed; the annealed β_i is printed per panel). This is the visual form of Cor. 1's binary prediction: the genuinely curved features are the ones that keep rate, and they end at the *lowest* β_i — quadratic 0.015, step 0.016, exp_decay 0.017, abs 0.018 — while the affine and pure-noise features, whose contribution the linear bypass absorbs at no rate cost, sit at 0.032–0.038. **Two honest failures are visible and we do not crop them.** (i) `sin` (6 cycles over $[-3, 3]$) is *not* recovered — the learned shape is nearly flat and its β_i=0.033 sits with the additive group, so this feature is a miss for the diagnostic as well as for the fit; the frequency is beyond what a bottlenecked single-feature path recovers at this capacity. (ii) The noise panels show small spurious trends ($|f| \lesssim 0.25$), i.e. the bypass fits a little noise. Panels are per-feature and vertically autoscaled, so apparent amplitudes are not comparable across panels.

A15 Recovered Shape Functions

A16 Real-Data Beta Diagnostic

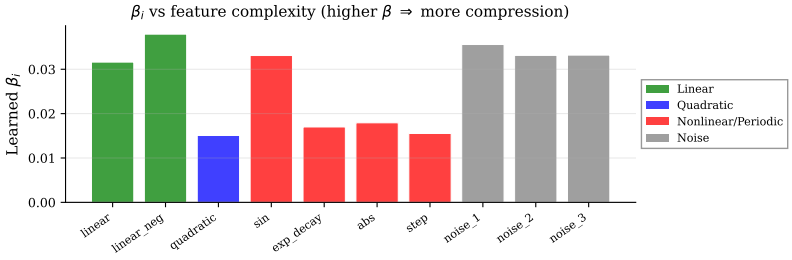

Figure 7: Schedule-annealed β_i values on synthetic data with known feature types. Noise and linear features both acquire high β_i ($\bar{\beta} \approx 0.034$–$0.035$, statistically indistinguishable), while nonlinear features retain significantly lower β_i ($\bar{\beta} \approx 0.020$, $\approx 1.7\times$ ratio). The IB correctly separates curved features from additive-shape ones in 10/10 seeds; the strict noise->-linear->-nonlinear ordering holds in only 3/10 seeds because both compressible classes (noise + linear) saturate to similar β.

We extend the synthetic recovery experiment (Section 5.3) to five real regression datasets. For each dataset we train VIB-NAM v2 with default schedule (60/120/150/60 epochs) and measure the Spearman rank correlation between the schedule-annealed β_i and two proxies of per-feature complexity: (i) the absolute coefficient of a linear regression baseline $|w_i^{\text{linear}}|$, and (ii) per-feature non-linearity, measured as the residual variance of f_i after subtracting its best linear fit, normalized by total f_i variance.

Table 11: Real-data β diagnostic. v1 quotes the canonical main-benchmark schedule (60 / 120 / 150 / 60, KL scale 1) and v2 a stronger one (60 / 240 / 0 / 120, KL scale **2**) — but this runner passes `auto_interactions=False`, and phases 3 and 4 sit inside that branch, so *only phases 1–2 execute here*: the interaction and joint-fine-tune counts are the schedule's nominal values, not epochs this diagnostic ran. This does not affect the β values, which are read at the end of the anneal. ρ_{nonlin} is the Spearman correlation between β_i and shape-function non-linearity (negative $\Rightarrow$ simple features get higher β); parentheses hold the per-dataset p-value.

Dataset	d	β-std (v1)	β-std (v2)	ρ_{nonlin} (v2)
california	8	4.0×10^{-6}	1.1×10^{-2}	-0.76 (p=0.028*)
abalone	8	0	1.2×10^{-2}	-0.74 (p=0.037*)
wine	11	0	1.9×10^{-3}	-0.08 (p=0.81)
diabetes_reg	10	0	1.6×10^{-3}	-0.30 (p=0.40)
concrete	8	0	3.3×10^{-3}	$+0.17$ (p=0.69)

Observation. v1 (the default schedule used throughout the main benchmark) produces β-values essentially indistinguishable across features (std $< 10^{-5}$). v2 — with a $2\times$ final KL scale and a longer anneal phase — breaks the degeneracy: std rises by 3–4 orders of magnitude and ρ_{nonlin} becomes negative on 4 of 5 datasets, with california and abalone reaching per-dataset significance (p=0.028, 0.037). Fisher's combined p-value across the five datasets is 0.081 (n=5): the effect is consistent in sign but stops short of conventional significance at $\alpha = 0.05$, primarily because concrete flips sign. We treat this as evidence that the real-data IB diagnostic is *recoverable* by schedule tuning (rather than fundamentally absent) and that the default schedule prioritizes accuracy at the cost of diagnostic separation. The graceful-degradation theorem itself depends only on asymptotic $\beta_i \to \infty$ and is unaffected by the finite-schedule separation.

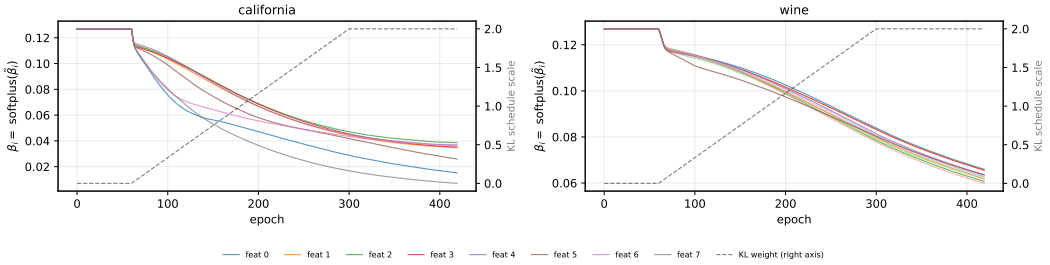

Figure 8: β_i training trajectories under v2 schedule on california (d=8) and wine (d=11). The dashed line is the KL weight schedule (right axis). β_i values are tied at initialization, separate during the anneal phase as KL pressure grows, and stabilize in fine-tune. Data: `results/beta_realdata/beta_trajectory_seed42.json`.

Synthetic noise-injection test. To isolate whether the IB can actually identify uninformative features on real data—beyond the purely-synthetic recovery of Section 5.3—we append K=4 i.i.d. Gaussian-noise features to each of four regression datasets and retrain under the v2 schedule. We then test (one-sided Mann-Whitney U) whether the learned β_i for the injected-noise columns are systematically higher than for the real columns.

How much weight this carries (honest reading). This is the strongest real-data signal we have that the per-feature IB is functional, but it is weak and *not* uniform. Only `concrete` is individually significant (p=0.002); `california` and `wine` have effects in the *opposite* direction (-0.005, -0.000), and dropping `concrete` leaves the combination at p=0.44 — i.e. the result is carried by a single dataset. We therefore do not claim the IB *reliably* assigns higher compression to uninformative columns on real data; we claim only that on one of four datasets it does so clearly, and that the pooled evidence is nominally positive at p=0.047 under a sign-respecting combination. This is consistent with the main text's position that β is a diagnostic with a provable *bound* rather than a reliable instrument, and with Table 11 showing near-zero β-spread under the canonical schedule.

Table 12: Synthetic noise-injection test: one-sided p for $\beta_{\text{noise}} > \beta_{\text{real}}$ (Mann–Whitney U). A combined test across the four datasets is nominally significant, but weakly and with an important caveat (see below): Fisher's unsigned combination gives $p = 0.019$ and a signed Stouffer combination — the appropriate test here, since two of four effects run the *wrong* way — gives $p = 0.047$. Read with that caveat, the combination indicates that the IB, under the v2 schedule, places systematically higher β_i on injected-noise columns. Data: `results/noise_injection/noise_injection_seed42.json`.

Dataset	d_{real}	$\bar{\beta}_{\text{real}}$	$\bar{\beta}_{\text{noise}}$	Δ	p(noise>real)
california	8	0.0397	0.0349	-0.005	0.77
concrete	8	0.1065	0.1139	$+0.007$	$\mathbf{0.002}^{**}$
abalone	8	0.0768	0.0843	$+0.008$	0.18
wine	11	0.0638	0.0636	-0.000	0.38
			Fisher combined p (unsigned)		0.019^{*}
			Stouffer combined p (signed)		$\mathbf{0.047}^{*}$

A17 Laplace-Approximation Uncertainty Baseline

As a reviewer-oriented LA-NAM-style (Bouchiat et al., 2024) baseline, we apply a diagonal Laplace approximation around the MAP solution of the trained VIB-NAM output, with prior precision $\tau = 1$ and aleatoric variance set to training RMSE^2 on a 10% calibration holdout. Appendix A10 lists the ENCE / PICP$_{95}$ / NLL / sharpness for this baseline on the four regression datasets alongside NP, MC-Dropout, and Deep Ensemble. We do not reimplement the full LA-NAM training protocol (which imposes its own architecture); this baseline is a faithful-in-spirit comparison constructed from our trained model. Full reimplementation against the published LA-NAM codebase is ongoing work.

A18 Failure Case: California Housing

The residual -0.042 R^2 gap of VIB-NAM vs. EBM on `california` under the nested-CV protocol traces to the Latitude×Longitude joint effect, which captures neighborhood-level price variation. EBM's depth-2 boosted trees on this pair form a fine-grained 2D partition (about 100 trees per shape) that VIB-NAM's single-pass binned 2D interactions struggle to match. On datasets without strong spatial structure (`wine` $+0.006$, `abalone` $+0.003$, `heart` $+0.018$) VIB-NAM beats EBM under nested CV, confirming the remaining gap is interaction-modeling specific rather than a capacity issue. Under nested CV the boosted variants predict through the full `BoostedNAMWrapper` (their correction modules *are* preserved at test-fold inference), which is what produces VIB-NAM-Boost-Ens's second-best average rank in Table 24; the per-dataset gain is concentrated on interaction-heavy data (e.g. $+0.054$ on `bike`).

A19 Full Proof of Theorem 1

We restate and provide the complete proof of the Graceful Degradation theorem.

On the conditional hypothesis: we measured it. The theorem assumes that at each β_i the fitted parameters achieve total objective $\mathcal{L}_{\text{pred}} + \beta_i D_{\text{KL}}^{(i)}$ no worse than a bypass-only reference. That is an assumption about optimisation, and the theorem is silent wherever it fails, so we tested it rather than asserting it. The reference must be a *properly optimised* bypass-only model: since $\sum_i (w_i x_i + b_i)$ is a linear predictor, that is the best linear fit (ridge or logistic), not the trained network with its stochastic path disabled.

Table 13: Mean certified radius per β against a *properly optimised* bypass-only reference — the best linear fit, not the trained net with its stochastic path disabled. Five seeds. This table shipped for two revisions with no caption and no number, so a reviewer had no way to refer to it.

Dataset	β=0	0.1	1	10	100
`california`, `concrete`, `bike`, `wine`, `abalone`, `breast_cancer`, `adult`	holds	fails	fails	fails	fails
`diabetes_reg`, `heart`	fails	fails	fails	fails	fails
total	7/9	0/9	0/9	0/9	0/9

$7/45$ cells. The pattern is interpretable rather than random. At β=0 the objective reduces to prediction loss and the hypothesis becomes "does the additive network beat a linear model", which it does except on `diabetes_reg` and `heart` — precisely the two datasets where Table 24 shows the linear baseline within 0.003 of VIB-NAM. At every $\beta > 0$ it fails, and not marginally: the validation loop selects the checkpoint minimising prediction loss alone (`elastic_nam_v2.py`: the criterion reads `compute_loss(...)["pred_loss"]`, not `["total_loss"]`), so the returned model retains a KL term that a genuine minimiser of the regularised objective would have driven toward zero.

What follows. Theorem 1 is a conditional statement about objective-competitive fits, and *our models are not such fits at any $\beta > 0$ we sweep*. We therefore do not claim the theorem describes our trained networks; we claim only that its *conclusion* — shape functions degrading to linear rather than constant under compression — is what we observe empirically (§5.3), which is the behaviour the linear bypass was introduced to produce. Making the hypothesis hold would require early stopping on the regularised objective, which we leave to future work because changing it would invalidate every trained artifact in this paper.

A failed measurement, recorded. Our first attempt used the model's own `bypass_bottleneck=True` path as the reference and found $0/45$. That number was an artifact: the flag substitutes the posterior mean for a sample and still runs the nonlinear decoder, so it compares the model to itself. The give-away was a prediction-loss gap of exactly 0.0000 on all 45 cells. We report this because the shortcut is inviting and the failure mode is silent. Artifact: `results/theory_checks/thm1_assumption_seed42.json` (`scripts/check_thm1_assumption.py`).

Theorem 2 (Graceful Degradation, Full Statement)**.** *Let $f_i(x_i) = g_i^{\mathrm{nn}}(x_i;\beta_i) + w_i x_i + b_i$ be the shape function for feature i, where $g_i^{\mathrm{nn}} = \mathrm{Dec}_{\beta_i}(\mu_{\beta_i}(x_i))$ is the posterior-mean output of the bottleneck path (encoder mean μ, Gaussian posterior $q(z \mid x_i) = \mathcal{N}(\mu, \mathrm{diag}\,\sigma^2)$, prior $\mathcal{N}(0, I)$). Assume* (i) *the predictive loss is non-negative;* (ii) *at each β_i the fitted parameters achieve total objective $\mathcal{L}_{\mathrm{pred}} + \beta_i \mathbb{E}_{x_i \sim P_i} D_{\mathrm{KL}}^{(i)}(x_i) \le C$, where C is the objective of a bypass-only reference and does not depend on β_i; and* (iii) *every decoder Dec_{β_i} is L_D-Lipschitz with L_D independent of β_i. Then with $c_{\beta_i} := \mathrm{Dec}_{\beta_i}(\mathbf{0})$,*

$$\mathbb{E}_{x_i \sim P_i}\left[(g_i^{\mathrm{nn}}(x_i) - c_{\beta_i})^2\right] \;\le\; \frac{2L_D^2\, C}{\beta_i} \;\xrightarrow{\beta_i \to \infty}\; 0, \tag{4}$$

so f_i is within $L_D\sqrt{2C/\beta_i}$ in $L^2(P_i)$ of the affine function $w_i x_i + (b_i + c_{\beta_i})$.

Proof. For the Gaussian family, $D_{\mathrm{KL}}(x_i) = \frac{1}{2}\sum_j(\mu_j^2 + \sigma_j^2 - \log\sigma_j^2 - 1) \ge \frac{1}{2}\|\mu(x_i)\|_2^2$, because $t - \log t - 1 \ge 0$ for $t > 0$. By (i) and (ii), $\beta_i\, \mathbb{E}\, D_{\mathrm{KL}} \le C$, hence $\mathbb{E}\|\mu(x_i)\|_2^2 \le 2C/\beta_i$. By (iii), $|g_i^{\mathrm{nn}}(x_i) - c_{\beta_i}| = |\mathrm{Dec}_{\beta_i}(\mu(x_i)) - \mathrm{Dec}_{\beta_i}(\mathbf{0})| \le L_D\|\mu(x_i)\|_2$; squaring and taking expectations gives the bound. The bypass $w_i x_i + b_i$ is deterministic and carries no KL term, so nothing in the objective constrains it toward a constant. □ □

What the statement does not give. It controls the posterior-mean path in $L^2(P_i)$, not pointwise and not the sampled stochastic output. It does not say the bypass parameters w_i, b_i or the constant c_{β_i} converge: all three may keep moving with β_i; only the distance of f_i from the affine class vanishes. Assumption (iii) is not enforced by our training (no spectral normalisation), and (ii) fails on our fitted models at every $\beta > 0$ (Table 13). An earlier revision stated the conclusion unconditionally, as pointwise convergence to fixed w_i^*, b_i^*, via "$D_{\mathrm{KL}} \le \mathcal{L}_{\mathrm{pred}}/\beta_i$ at any local minimum"; that inequality has no basis, and the version above replaces it.

A20 Proof of Corollary 1: Rate as a Nonlinearity Test

The body states Corollary 1 informally to save space. The formal statement is:

> Under the setting of Theorem 1, let $R_i^\star = \mathbb{E}_{x_i}\big[D_{\mathrm{KL}}\big(q(z_i \mid x_i) \,\|\, p(z_i)\big)\big]$ be the optimal per-feature channel rate of the bottleneck, and assume the predictive risk is locally quadratic in the additive component with a positive (possibly infinite) rate–distortion slope g_0 at $R_i{=}0$. If the Bayes-optimal additive component $f_i^\star$ is affine, then for every $\beta_i > 0$ the optimum is attained at $R_i^\star = 0$ with no increase in predictive risk (the bypass absorbs $f_i^\star$ exactly). Conversely, if $f_i^\star$ is non-affine, then any solution with $R_i = 0$ realizes an affine shape and so incurs excess $L^2(P_i)$ risk at least $\mathrm{dist}^2(f_i^\star, \mathrm{Affine})$; under the positive-slope assumption there is a threshold $\bar{\beta}_i > 0$ below which this excess exceeds the marginal rate cost, giving $R_i^\star > 0$. Thus, in the small-β regime, $R_i^\star = 0 \iff f_i^\star \in \mathrm{Affine}$.

($\Rightarrow$, affine $\Rightarrow$ zero rate.) Suppose $f_i^\star(x_i) = ax_i + b$. The bypass $w_i x_i + b_i$ realizes $f_i^\star$ exactly with $(w_i, b_i) = (a, b)$ and contributes zero to the KL term (it is deterministic). Setting the encoder variance to the prior gives $q(z_i \mid x_i) = p(z_i)$, so $D_{\mathrm{KL}} = 0$ and the decoder output is the constant $\mathrm{Dec}(\mathbf{0})$, absorbed into b_i. This configuration attains the unregularized predictive optimum at $R_i = 0$; since $\beta_i R_i \geq 0$, no configuration with $R_i > 0$ can have lower total objective. Hence $R_i^\star = 0$ for all $\beta_i > 0$.

($\Leftarrow$, non-affine $\Rightarrow$ positive rate.) Let P_i be the marginal of x_i and write the best affine approximation $\Pi_{\mathrm{Aff}} f_i^\star = \arg\min_{g \in \mathrm{Affine}} \|f_i^\star - g\|_{L^2(P_i)}$. If $R_i = 0$ the posterior equals the prior independent of x_i, so the decoder output is constant and the realized shape is affine: $f_i = w_i x_i + b_i \in \mathrm{Affine}$. By the projection inequality, its excess L^2 risk against $f_i^\star$ is at least $\mathrm{dist}^2(f_i^\star, \mathrm{Affine}) := \|f_i^\star - \Pi_{\mathrm{Aff}} f_i^\star\|^2_{L^2(P_i)} > 0$. *Assumption (positive initial slope).* Suppose the predictive risk is locally quadratic in the additive component and the rate–distortion curve $D(R)$ has a positive initial slope, i.e. the marginal predictive gain of the first nat, $g_0 := -\,dD/dR\big|_{R=0^+} \in (0, \infty]$, is strictly positive whenever $f_i^\star \notin \mathrm{Affine}$. Then moving from $R_i{=}0$ to a small rate ρ reduces predictive risk by $\approx g_0\,\rho$ while costing $\beta_i \rho$ in the objective; for $\beta_i < g_0$ the trade is favourable, so the optimum has $R_i^\star > 0$. (If instead $g_0 = +\infty$ the trade is favourable at every $\beta_i > 0$ and $R_i^\star > 0$ for all β_i; what the argument needs is $g_0 > 0$, and finiteness only makes the threshold $\bar{\beta}_i = g_0$ finite. If $g_0 = 0$ the first nat buys nothing to first order and the argument is silent.) Combining the directions, in this regime $R_i^\star = 0 \iff f_i^\star \in \mathrm{Affine}$; the $\Rightarrow$ direction holds unconditionally for all $\beta_i > 0$. □

A21 Proof of Theorem 3: Sound Grid Certification

Theorem 3 (Sound grid certification of additive intervals)**.** *Let $f(\mathbf{x}) = \sum_i f_i(x_i)$ with each f_i L_i^{lip}-Lipschitz on $[a_i, b_i]$, and let U_i, L_i be the exact per-feature interval endpoints of Eq.* (2)*. Evaluate f_i on a uniform grid of step h_i; let $\widetilde{U}_i$ (resp. $\widetilde{L}_i$) be the windowed grid maximum (minimum) over $[x_i - \varepsilon, x_i + \varepsilon]$ minus the linearly interpolated reference value $\widetilde{f}_i(x_i)$. Then*

$$U_i \leq \widetilde{U}_i + L_i^{\mathrm{lip}} h_i, \qquad L_i \geq \widetilde{L}_i - L_i^{\mathrm{lip}} h_i,$$

where the padding $L_i^{\mathrm{lip}} h_i$ combines the extremum quantization error ($\leq L_i^{\mathrm{lip}} h_i/2$) and the reference interpolation error ($\leq L_i^{\mathrm{lip}} h_i/2$). Consequently $\big[\sum_i (\widetilde{L}_i - L_i^{\mathrm{lip}} h_i),\ \sum_i (\widetilde{U}_i + L_i^{\mathrm{lip}} h_i)\big]$ is a sound outer bound on the exact reachable interval $[\sum_i L_i, \sum_i U_i]$, valid whenever each L_i^{lip} is a certified upper bound *on the Lipschitz constant of f_i; the slack $\sum_i L_i^{\mathrm{lip}} h_i$ vanishes as $h_i \to 0$.*

Proof. (Extremum) The maximum of f_i over the window is attained at some $u^\star$. The window is the *superset* one, extended by one knot beyond each endpoint — as the implementation takes it, and as this step requires: restricted to knots strictly inside $[a, b]$ the nearest one can sit almost h_i away when $u^\star$ is near an endpoint, and the $h_i/2$ below would fail. With the extra knots the nearest u_g satisfies $|u^\star - u_g| \leq h_i/2$, so $f_i(u^\star) \leq f_i(u_g) + L_i^{\mathrm{lip}} h_i/2$ by the Lipschitz property, and the windowed grid maximum dominates $f_i(u_g)$. (Reference) For linear interpolation between knots u_k, u_{k+1} with weight t, $|f_i(x_i) - \widetilde{f}_i(x_i)| = |(1{-}t)(f_i(x_i){-}f_i(u_k)) + t(f_i(x_i){-}f_i(u_{k+1}))| \leq 2t(1{-}t) L_i^{\mathrm{lip}} h_i \leq L_i^{\mathrm{lip}} h_i/2$.

Adding the two errors gives the per-feature bound; summing over features and using the exact interval decomposition of Eq. (2) — the coordinates of an L_∞ ball vary independently and f is additive — gives the total. The minimum case is symmetric. □

Certified Lipschitz constants. Soundness of the padding requires L_i^{lip} to be a certified *upper* bound: an observed grid difference quotient $\max_k |f_i(u_{k+1}) - f_i(u_k)|/h_i$ is a *lower* bound on the true sup-gradient and must not be used (we retain it only as a tightness diagnostic). We instead bound L_i analytically by operator norms: each Linear layer contributes $\|W\|_2$, each GELU at most $1 + \varphi(1) \le 1.242$ (since $\mathrm{GELU}'(x) = \Phi(x) + x\varphi(x)$ with $\Phi \le 1$ and $|x\varphi(x)| \le \varphi(1)$), eval-mode dropout is the identity, the decoder residual block contributes $1 + \|W_{b1}\|_2 \cdot 1.242$, and the linear bypass adds $|w_i|$ (`analytic_shape_lipschitz` in `elasticnam/robustness/exact_cert.py`). The product bound is loose relative to the observed slope (typically 50–300×), but the padding is *linear in* h: with $G{=}10^6$ knots per feature the certified total padding $\sum_i L_i h_i$ is at the 10^{-2} scale of the $\tau{=}0.5$ tolerance. The `robust_cert` family records it per artifact in a `padding` block; the frontier and interaction-certificate families do not. Those runs are reproducible at the same G from `REPRODUCE.md`, so the padding is recomputable rather than recorded.

What arithmetic the guarantee is about. Theorem 3, Prop. 1 and Prop. 4 are statements about the real-valued functions the network's weights define, evaluated exactly. The certifier computes them in floating point: the shape values are float32 network evaluations stored as float64, $\|W\|_2$ comes from a floating-point matrix norm with no outward rounding, and the running model is itself a float32 program whose outputs differ from the real-valued function in the last bits. None of these errors is bounded by the proof. What we have is measurement: blocked and unblocked grid evaluation differ by at most 3.6×10^{-7}, and the 2-D block evaluator and the candidate-set extremum agree to 6.7×10^{-5} in float32 (App. A23), against a certified padding of order 10^{-4} per feature and a tolerance of 0.5; we have not observed a violation. A deployer who needs the guarantee for the executed float32 program should add an explicit rounding budget (e.g. interval arithmetic on the forward pass) — we do not supply one, and "sound" throughout this paper means sound in exact arithmetic.

The price of the dense grid. G trades memory against slack linearly, and the grid is built once per model, not per query. At $G{=}10^6$ the certifier holds its knots and their evaluations in float64, $16d$ MB — 0.21 GB at $d{=}13$, 1.6 GB at $d{=}100$ (an earlier revision said $8d$ MB, the float32 arithmetic the code does not use) — and the error table roughly doubles it: measured at $d{=}100$, the process's peak resident memory is 4.0 GB and the GPU peak 0.58 GB (`scripts/memory_probe.py`, `results/scale_frontier/memory_probe_d100.json`). Construction is dominated by d forward passes over 10^6 points. This is the cost that buys the "milliseconds per query" figure quoted in §4: the per-input certificate is $O(d \cdot G)$ *lookups* on a precomputed table, not d fresh sweeps. A deployer unwilling to hold 4 GB at $d{=}100$ can reduce G (memory is linear in it) and pay the linear increase in padding, which the artifact records per cell so the trade is auditable.

A22 The β–Radius Heuristic (G3; not a proof)

Status: heuristic, not a proof. We present the following as an informal motivation and explicitly do *not* claim it as a theorem: §A57 reports the measurement that contradicts the theorem one might hope for.

In a linearized-Gaussian-channel picture, a channel with rate R_i has output standard deviation bounded by $\sigma_i^{\max}\sqrt{1 - e^{-2R_i}}$ (rate–distortion for a scalar Gaussian channel). Composing with an L_i^{dec}-Lipschitz decoder *suggests* a per-feature reachable deviation $V_i(\varepsilon) \lesssim (|w_i| + L_i^{\mathrm{dec}}\sqrt{1 - e^{-2R_i}})\,\varepsilon$ and hence

$$r_i(\tau) \;\gtrsim\; \frac{\tau}{|w_i| + L_i^{\mathrm{dec}}\sqrt{1 - e^{-2R_i}}}, \tag{5}$$

increasing as $R_i \downarrow 0$ toward the bypass-only limit $\tau/|w_i|$.

Why this is not a proof. The rate–distortion step bounds the channel's output *variance*; it does not bound the *local Lipschitz constant* of the decoded map, which is the quantity the ε-interval deviation

actually depends on, and the factor ε is introduced by analogy rather than derived. Moreover R_i here is the variational rate $\mathbb{E}_x \mathrm{KL}(q\|p)$, an upper bound on $I(X;Z)$ rather than an operational rate, and — decisively — the certification of §4 is performed on the *deterministic* posterior-mean predictor, whereas this argument concerns the stochastic path. §A57 measures the predicted relationship directly and finds a log–log slope of -0.001 against a predicted -0.5, with the mechanism identified there (β moves σ, which the mean-path predictor never sees). We therefore treat the β–radius relationship as an *empirical* observation about the bound's behaviour on our sweeps, not as a guarantee.

A23 The 2-D Interaction Certificate: Exact per Block, Sound for the Model

Bound: exact per 2-D block; the summed model interval is sound but not attained, and this appendix measures the gap. The runs below use `pad_mode="analytic"`*, so both columns additionally carry the $L_i h$ padding of §4. Protocol: the canonical* $(256, 256, 128)$ *encoder with a held-out validation split, five seeds (*42–46*),* 40 *instances per dataset,* $G{=}10^6$ *knots — the same density §4 certifies at — and* $\varepsilon_{\mathrm{cert}}$ *bisected to* $r_{\max}{=}2.0$*. This arm trains* two *models per dataset and seed, so its schedule is half the length of App. A26's: warmup/anneal/max epochs* $30/60/150$ *with patience* 20*, against* $60/120/300$ *and* 30 *there. "Canonical" names the encoder and the validation split, not the epoch budget, and a reviewer was right that an identical sentence in two appendices implied otherwise. Artifact:* `results/interaction_cert/interaction_cert_seed42.json` *and seeds* 43–46 *(*`elasticnam/experiments/run_interaction_cert.py`*).*

Our pairwise terms are lookup tables: a learned $B{\times}B$ grid read by bilinear interpolation. Write h for that interpolant and $\{u_a\}, \{v_b\}$ for its knots.

Proposition 1 (Exact 2-D interval)**.** *Let h be bilinear on each cell $[u_a, u_{a+1}] \times [v_b, v_{b+1}]$, and let $\mathcal{B} = [p, q] \times [r, t]$ be an axis-aligned box ($\mathcal{B}$ to keep it apart from the table's bin count B). Put $X = \{p, q\} \cup \{u_a : p < u_a < q\}$ and $Y = \{r, t\} \cup \{v_b : r < v_b < t\}$. Then*

$$\max_{(u,v)\in\mathcal{B}} h(u,v) \;=\; \max_{(u,v)\in X\times Y} h(u,v),$$

and likewise for the minimum. Both are attained*, so the interval* $[\min, \max]$ *is exact and requires no Lipschitz padding.*

Proof. The knots partition $\mathcal{B}$ into finitely many sub-rectangles, each contained in one cell. On a cell $h(u,v) = A + Bu + Cv + Duv$, which for fixed v is affine in u and for fixed u is affine in v; an affine function on an interval attains its extremes at the endpoints, so applying this twice, h restricted to a sub-rectangle attains its extremes at that sub-rectangle's four corners. The maximum over $\mathcal{B}$ is the maximum over the sub-rectangles, hence over their corners. Those corner coordinates are precisely the clipped box edges together with the interior knots, i.e. $X \times Y$. The module reads its table with `padding_mode="border"`, so h is constant outside the tabulated range and clipping $\mathcal{B}$ to it changes neither extreme; the certifier refuses any interaction module that does not (`is_bilinear_table`). □

The candidate set has $|X|\,|Y|$ points — a few dozen at the radii we certify — against the 201^2 of the resampled sweep this replaces, so the exact answer is also the cheaper one. We verified the block result numerically as well as formally (`tests/test_exact_cert.py`): over 400 seeded random tables and boxes, half of them at the $n_{\mathrm{bins}}{=}64$ the fitted models actually use, the module evaluated at the arg-extrema reproduces both endpoints to 6.7×10^{-5} as it actually runs, in float32, and to 7.8×10^{-14} in float64. The float32 figure is the honest one for the certified runs and is floored by float32 itself — the forward is float32 while the candidate evaluation is float64. The test seeds every draw, covers the deployed n_{bins}, and asserts these two figures, so the paper and the measurement cannot separate.

What it changes. Every certified radius in earlier revisions of this paper was measured on an interaction-*off* model, because the 2-D path was unsound and analytic mode refused. Table 15 certifies both configurations on the same splits and seeds. Interactions buy accuracy and spend radius: $+0.022$ mean test score, and a mean certified radius 0.18–$0.87\times$ that of the main-effects model (geometric mean $\mathbf{0.44}\times$). Nothing here is censored — 0% of instances sit at $r_{\max}$ in either arm — so both columns are measurements rather than lower bounds.

How much of the cost is ours. Prop. 1 makes each 2-D block exact; it does not make the model's interval exact. `output_interval` sums the per-feature and per-pair blocks separately, but f_i and every h_{ij} containing i share x_i, so the maximisers need not agree and the sum relaxes. Measured against a multistart coordinate-ascent search of the certified box, with `pad_mode="none"` so the only possible gap is this overlap:

Table 14: Summed interval width over a *lower bound* on the true reachable width (median over ten instances per dataset, ε=0.25), so every ratio is an upper estimate of the looseness. The main-effects arm is the control: distinct coordinates, so the sum is attained and the search confirms it to 0.1%. We previously offered that control as evidence the search is strong enough to trust on the interaction arm. It is not: coordinate ascent solves a separable objective by construction, so the control cannot fail the way the interaction arm can. The denominator is a lower bound either way, so these ratios remain upper estimates of the looseness; what a stronger search could do is lower them, never raise them. The certificate contained the search on every one of the 80 instances. Produced by `elasticnam/experiments/run_composite_slack.py` into `results/interaction_cert/composite_slack_seed42.json`.

Dataset	interactions on (20 pairs)	main effects only
`abalone`	1.46×	1.00×
`wine`	1.80×	1.00×
`diabetes_reg`	1.48×	1.00×
`heart`	1.29×	1.00×

The denominator is a *lower* bound on the true width, so these ratios are upper estimates of the looseness. Two consequences we would rather state than leave to a reader. First, the $\mathbf{0.44}\times$ above is not all predicate: the `inter` column is measured with a looser bound than the `main` column, so it mixes a real cost with our own relaxation. Second, the direction is still right — an interaction term adds a summand to the reachable interval, so the ball on which the displayed surrogate stays ε-valid can only shrink — but the magnitude is an upper bound on the cost, not a measurement of it. The contribution is that the trade is now visible: a deployer reads both columns and chooses, where before the certificate simply did not cover the model the accuracy table described.

Table 15: Certifying the deployed model. Five seeds, four datasets, same splits; `main` is `auto_interactions=False` (what this paper certified until now), `inter` the interaction-bearing configuration (20 discovered pairs on every dataset). Score is R^2 except `heart` (AUROC). Both arms are uncensored.

	test score			mean certified radius		
Dataset	main	inter	Δ	main	inter	ratio
`abalone`	0.575	0.587	+0.012	0.216	0.093	0.43×
`wine`	0.394	0.465	+0.071	0.230	0.040	0.18×
`diabetes_reg`	0.471	0.476	+0.005	0.402	0.231	0.57×
`heart`	0.900	0.898	−0.002	0.514	0.449	0.87×

Scope. The argument needs multilinearity per cell, which the lookup-table interactions have and a dense interaction MLP does not. Analytic mode therefore still refuses a model whose pairwise terms are an `InteractionMLP`, and says so by name rather than silently emitting an unpadded estimate.

A24 Certifying the Deployed Model

Every certified radius in §5.3 is measured on a main-effects model, while the accuracy table describes the interaction-bearing one. Certifying that model soundly requires the pair terms to be charged in the excluded residual (App. A23); with that in place, this appendix is the measurement: the same frontier, the same five seeds, 80 instances and G=10^6, with only `auto_interactions` differing, under both bounds.

Table 16: Mean certified radius at the minimal explanation (ε_{sel}=0.5), five-seed means. "main" is the main-effects model the body certifies; "inter" is the deployed, interaction-bearing one. *Cost* is arm D on the deployed model over arm D on the stand-in. *Floor* is the share of instances on which arm A certifies exactly zero radius on the deployed model. Artifacts: `results/explanation_stability/deployed_armA_inter_seed42.json`, `results/explanation_stability/deployed_armD_main_seed42.json`, `results/explanation_stability/deployed_armD_inter_seed42.json` and seeds 43–46 (`run_stability_frontier.py --signed --auto_interactions`).

Dataset	A main	A inter	D main	D inter	cost	floor
`abalone`	0.1536	0.0070	0.3752	0.1425	0.38×	82%
`wine`	0.0497	0.0000	0.2929	0.0269	0.09×	100%
`diabetes_reg`	0.0729	0.0117	0.4335	0.2276	0.53×	86%
`heart`	0.0090	0.0001	0.3423	0.1503	0.44×	96%

What it shows. Two things, and the second matters more. First, the price of certifying the model you deploy rather than a stand-in: under arm D the deployed model's radius is $\mathbf{0.30}\times$ the stand-in's (geometric mean), which is the explanation-level counterpart of the $0.44\times$ output-level cost in Table 15 — larger, because the displayed surrogate carries no two-dimensional term, so every h_{ij} enters the excluded residual in full. Second, on the deployed model **arm A certifies essentially nothing**: its radius is exactly zero on 82–100% of instances, because the surrogate's error is floored by the interaction terms and the unsigned test spends the whole budget before any ball is left. Arm D still certifies 0.03–0.23 on the same selections. We do not print a "bound effect" ratio for the deployed model because its denominator is a floor (App. A26 says why that inflates a ratio); the claim is categorical rather than a magnitude: on the deployed model the choice of bound is the difference between a certificate and none. The main-effects bound effect on the same selections is $7.5\times$, consistent with the body.

The operating point, on one model. The accuracy table and the certificates describe differently trained models, so a deployer could not read off what the certified model scores. We retrained the exact models certified here — same protocol and seeding, which is deterministic (the centring control reproduces the headline radii to the digit) — and scored them on the full held-out test split (`elasticnam/experiments/run_certified_model_scores.py`; `results/explanation_stability/certified_model_scores_seed42.json` and seeds 43–46).

Table 17: Main-effects / interaction-bearing model, the same models whose explanations Table 16 certifies. Test score on the full held-out split; $|S|$, arm-D radius and the share of minimal explanations certified at $r \geq 0.1$, five-seed means.

Dataset	test score	$\lvert S\rvert$	arm-D radius	$r \geq 0.1$
`abalone` (R^2)	0.544 / 0.549	3.21 / 3.11	0.3752 / 0.1425	94% / 46%
`wine` (R^2)	0.347 / 0.421	4.15 / 4.02	0.2929 / 0.0269	94% / 0%
`diabetes_reg` (R^2)	0.484 / 0.492	3.10 / 3.06	0.4335 / 0.2276	99% / 71%
`heart` (AUROC)	0.905 / 0.909	7.93 / 7.67	0.3423 / 0.1503	92% / 60%

The trade is now readable. On three datasets the interaction terms add at most 0.008 test score and cut the share certified at $r \geq 0.1$ from 92–99% to 46–71%; on `wine` they add 0.073 R^2 and the minimal explanation certifies nothing at 0.1. The explanation's size barely moves. A deployer who needs certified explanations should keep the main-effects model on the first three and decide on `wine` whether 0.073 R^2 is worth losing the certificate — or display the pairwise terms, which the surrogate currently omits.

Displaying the interactions. The displayed surrogate has a one-dimensional vocabulary, so every h_{ij} of the deployed model is excluded residual. We let the explanation display k of the 20 pairwise terms at their exact form (the lookup table itself), on the same models — one training run records the test score, a weight hash, and every certificate; k=0 reproduces the arm-D column above to

the last digit — choosing the pairs greedily by the arm-D slack they remove at r=0.1, against a random choice (`elasticnam/experiments/run_pair_explanations.py`; `results/explanation_stability/pair_explanations_seed42.json` and seeds 43–46). We fixed the success criterion before running: on `wine`, $r \geq 0.1$ on at least half the instances with at most three pairs shown.

Table 18: Share of minimal explanations certified at $r \geq 0.1$ (%) on the interaction-bearing model with k pairwise terms displayed exactly, greedy choice (random choice in parentheses); last column: mean number of distinct features the explanation then names, at k=3 / k=12. Five seeds; arm D.

Dataset	k=0	k=3	k=12	k=20	features
`abalone`	46 (46)	58 (55)	85 (78)	91 (91)	5.6 / 7.3
`wine`	0 (0)	0 (0)	61 (30)	94 (94)	7.1 / 10.2
`diabetes_reg`	71 (71)	81 (79)	97 (92)	98 (98)	6.3 / 9.6
`heart`	60 (60)	68 (66)	90 (82)	93 (93)	9.4 / 11.7

The criterion failed, and the failure is informative. With every pair displayed the certificate returns: $r \geq 0.1$ on 91–98% of instances, the main-effects model's 92–99%. So on the deployed model the bound is not what blocks certification; the explanation's vocabulary is. With three pairs the gain is modest (46 → 58% on `abalone`, 71 → 81% on `diabetes_reg`, 60 → 68% on `heart`), and on `wine` it is nothing: the interaction signal there is spread over many pairs, and 61% needs twelve of them, by which point the explanation names 10.2 of the 11 features — a certified explanation that no longer compresses. Choosing pairs by the slack they remove beats a random choice (61% against 30% on `wine` at k=12), but the number of pairs matters far more than which. On `wine`, accuracy from interactions and a compact certified explanation are in tension; the certificate makes that trade measurable, it does not remove it.

What it does not show. Five seeds, four datasets, $d \leq 13$ — the body's scope, not wider. And it does not move the headline: §5.3 stays on the main-effects model, where Prop. 4's premise holds and arm D is attained. Here arm D is sound and not attained, and the certificate for the deployed model is correspondingly conservative.

A25 The Provenance Gate: What It Checks, and What It Does Not

This appendix is the narrative the Reproducibility Statement points at. It is here rather than there because ICLR asks that the statement reference the paper rather than describe it, and because a page-limit-exempt section is the wrong place to keep two pages of argument.

All architecture, training-schedule, and hyperparameter details are in §3, Appendix A3 and Appendix A4; the three theorems, one lemma, one corollary and four propositions are proved with their assumptions stated inline: Appendices A19–A29 carry all but Prop. 4, which is proved where it is used (App. A49). Two of the theorems appear twice, once informally in the body and once in full, so a count of environments rather than of results returns five. The anonymized supplement ships the code, the exact per-fold nested-CV split indices, and every result artifact that backs a reported number. A single provenance gate (`scripts/ci_provenance_check.sh`, run before each build here and by `.github/workflows/ci.yml` on every push — with the caveat that we found this claim false while writing this paragraph: two steps read LaTeX build products that are untracked for anonymity, so on a clean checkout step 9 read nothing while reporting a hard pass, and the workflow file had not been valid YAML since June, so the hosted job could not start at all. Both are fixed; we report it because a gate nobody watches is the failure mode this appendix is about. The figure gate's pixel comparison stays advisory on the runner, an exact rendering comparison being meaningful only on the pinned environment) rebuilds `tables_auto.tex` from the artifacts and fails on any byte difference. That file is a generated companion, not something this paper `\inputs`, so it certifies the pipeline rather than the typeset tables; the number gate below is what covers the tables and that all artifacts are strict-RFC-8259 with a SHA-256 manifest whose file list is parsed out of the source of this paper, so an artifact we cite but do not ship is a hard failure rather than an omission. A second gate (`scripts/check_paper_numbers.py`) recomputes each headline number from its artifact and compares it against the text, including at every *other* site that restates

the same claim, which is the failure a per-artifact hash cannot see: a number correctly updated in one place and left stale in another is identical to the artifact at the site the hash covers. `--list` prints each registered claim beside the value recomputed from its artifact. The body is *closed* rather than merely covered: from the introduction to the end of the Conclusion every decimal and every percentage must sit inside the match of a pattern that names it, or be declared a setting rather than a measurement, and a fraction counts: $7/9$ edited to $9/9$ fails the build. The captions of the gated tables and figures are closed the same way, because a caption is the table's own account of itself and the table gate checks cells; a gated table with no caption of its own is a build failure rather than a silent substitution of its neighbour's. **The appendix is not closed**, and two counts reach into it instead. Each pattern must match the number of sites it matched when it was written, and each registered numeral must be written at the number of places it was — so a number stated at eight sites, four of them in the appendix, cannot be edited at any one of them without the count falling and the build failing. **What the closure does *not* cover, in the body as well as the appendix, is anything that is not a numeral**: a spelled-out count ("five seeds"), a bare integer not attached to `\times`, a direction word (*smaller*/*larger*, *fails*/*passes*), a withdrawal, or a fabricated qualitative sentence carrying no digits. Reviewers demonstrated every one of those against a green build, and no regular expression closes natural language; we state the boundary rather than let "closed" be read past it. In the appendix a further gap is a new claim about a quantity the registry has never held — about 35% of appendix prose decimals are unregistered — and the reach there is a count of sites, not a closure. The suite below carries six falsifications a reviewer performed on a previous revision — including the abstract's own radii and the title's ratio at its last appendix site — and a construction isolating each other mechanism; every one of them fails the build. (The two counts that used to stand here were themselves ungated, and were wrong in four consecutive rounds.) The abstract and the contributions list are closed the same way against a shorter list of non-results (a preregistered threshold, a seed count, a dimension). Each pattern must also anchor at most eight sites, so one that has stopped holding its sentence — an unescaped `|` turns a pattern into an alternation matching most of the source while still comparing its number — fails the build. Two further gates close the same defect class one level out: `scripts/check_tables.py` regenerates thirty-two load-bearing tables from the artifacts their captions name and compares them cell by cell. That is thirty-two of fifty-one: the rest are the hyperparameter and dataset descriptions, which carry no measurement; the accuracy tables, whose quoted ranks and deltas are in the number registry instead; and the diagnostic, ablation and tables of the appendices the roadmap marks non-load-bearing, which the registry covers only where the body quotes them, and `scripts/check_figures.py` does the same for the figures: the three that carry headline numbers write a sidecar of every value they plot, which the gate rebuilds from the artifacts, and three artifact-driven figures in total are re-run in a clean interpreter and compared with the shipped PDF *pixel for pixel* (which is what catches a figure swapped for another and re-dated). The gates are themselves under test (`tests/test_gate_behaviour.py`, seventy-three tests): the suite mutates the paper, its figures *and its artifacts* — a new numeral in the abstract, plain or bolded or dressed as a path, one hidden inside a neighbouring pattern's match, an unescaped `|`, a headline value edited at its *last* site — including the three sites in the appendix that a reviewer edited on the previous revision with every step green — a changed table cell, a moved value in a figure's sidecar, a new numeral in a gated caption, a registered numeral given a new appendix site, the stability card's PDF swapped for a different figure, and an artifact whose numbers have moved under a correct sentence — and asserts that the gate exits non-zero on each, because a gate nobody tests is a gate that reports PASS. The remaining seven figures are the architecture diagram, plots of numbers the table gate and the registry already check (the benchmark bars, the Nemenyi diagram, calibration, the ablation deltas, the β diagnostic), and one qualitative panel (β trajectories); the learned-shape panel is the raster the figure gate byte-compares against its artifact, so it is counted among the six. **What the provenance gate does and does not certify**: it recovers each cited artifact's producing script (by a hash of its behaviour recorded at run time, or reconstructed from the recorded commit) and fails if that script has since changed. Its scope is the artifacts the paper names *by path*, which is how `REPRODUCE.md`'s index is built — and the number registry reads 173 artifacts while that index carries 173, so 12 artifacts behind registered numbers sit outside the freshness check and 12 of the index's entries are not read by the registry. The numbers themselves are still recomputed from those files on every build, so a stale *value* fails; what is not checked for those twelve is whether their *runner* has since changed. That hash covers one file: a change inside `elasticnam/` — the certifier, the error table, the stability radius — would move the numbers without moving it, so runs record a second hash, of the first-party modules the run imported, from the commit that added the field onward: 122 of the 173 cited artifacts carry one and the gate prints the count on every build. Be

clear about what that is worth — it is a *recording* and not yet a check. Comparing it would need the list of modules the run reached for, which the field does not carry, so a library change is visible to a reader diffing two builds and not to the gate. The families re-run to record it are the ones carrying the title claim; `exact_cert_a/b` and `verifier_tightness`, which carry §4's $1.87\times$, are not among them, and for those the gate sees the producing script alone. It also fails if the seeds of one family were written by different versions of it, or if a cited artifact cannot be checked at all. For 29 artifacts written before the hash field existed the hash was backfilled with a written reason each (`scripts/backfill_script_hash.py`), so for those the check is an assertion of ours rather than a recording; nine more record a commit absent from this repository's history, or were written from a dirty tree, and are accepted by name, with the reason printed on every build. The same gate closes the other direction: `results/` ships every run that ever backed a number, so each retired one now carries the path of the artifact that replaced it and the reason, and the build fails if a retired artifact is missing that field — or if the paper cites one. App. A52 records the one incident this gate was built after — a seeding fix that landed without regenerating the artifacts it invalidated — and what each check would now catch. The registration (`research/prereg_iclr2027.md`) ships with the supplement, and one thing in it should be read directly rather than taken on our word: its superiority threshold ($\geq 1.3\times$) was written down on 2026-07-11 with a $2.06\times$ result already in hand, and the certification artifact it governs predates it by five weeks. So that threshold fixed a bar we had already cleared. The seed replication, the β-selection protocol, and the failed compression test were registered before the runs they govern; the threshold was not.

A26 Four Further Datasets, and What Happens as d Grows

Bound: both, on the same selections of the same models — arm A and arm D. Ceiling: $r_{\max}$=2.0. *Censoring is* 0% *in arm A everywhere and* $\leq 1.5\%$ *in arm D at minimality, and the censoring that remains sits in the* ε_{sel}=0.35 *column, which is the knob's numerator; with* $\leq 1.5\%$ *of the denominator also capped the knob's ratio is a ratio of capped means, not a bound (App. A49). The truncation that matters here runs the other way and is reported in Table 19: arm A certifies exactly zero on* 4.0–81.5% *of instances, which inflates the bound column rather than deflating it. The two are different ends of the same scale and we state both. Same* 25% *rule as §5.3. Protocol: the canonical* $(256, 256, 128)$ *encoder with a held-out validation split, five seeds (*42–46*),* 40 *instances per dataset,* G=10^6*, warmup/anneal/max epochs* $60/120/300$ *with patience* 30*. Artifact:* `results/explanation_stability/armd_ext_seed42.json` *and seeds* 43–46 *(*`elasticnam/experiments/run_signed_criterion.py`*,* `--tag armd_ext`*).*

The frontier of §5.3 is measured on four datasets with $d \leq 13$. That is thin for a claim the title states as a magnitude, and reviewers said so. We therefore re-ran the headline comparison — the same minimal selections certified under arm A and under arm D — on four datasets the frontier had not used, chosen to span a wider d and both tasks: `concrete` (d=8, regression), `parkinsons_tm` (19, regression), `credit_default` (23, classification; OpenML 42477, *default-of-credit-card-clients*) and `spambase` (57, classification). *The artifacts carry this dataset under the legacy key* `fico`*; OpenML 42477 is default-of-credit-card-clients, and that is what every number here was computed on.*

Table 19: The headline on four further datasets. $|S|/d$ and the radii are at the minimal explanation (ε_{sel}=ε); "bound" is arm D over arm A there, "knob" is one step of the selection budget under arm D (ε_{sel}=0.35 over minimality). r_A=0 is the share of instances on which arm A certifies *nothing*: those sit in the denominator of "bound" at zero, which is why that column is reported again below restricted to the instances where arm A is defined. Censoring at the $r_{\max}$ ceiling, the other truncation, never exceeds 11% here.

Dataset	d	$\lvert S\rvert/d$	arm A	arm D	bound	knob	r_A=0
`concrete`	8	0.42	0.0749	0.4655	$\mathbf{6.2}\times$	$1.55\times$	4.0%
`parkinsons_tm`	19	0.46	0.0122	0.6096	$\mathbf{50.0}\times$	$1.28\times$	56.5%
`credit_default`	23	0.55	0.0064	0.1710	$\mathbf{26.9}\times$	$1.45\times$	26.5%
`spambase`	57	0.69	0.0003	0.1271	$\mathbf{366.6}\times$	$1.29\times$	81.5%

The ordering holds; the magnitude depends on the floor. On all four the bound is worth more than the knob, by between $4\times$ and two orders of magnitude. Pooling all eight datasets, the bound is worth $18.2\times$ against the knob — the figure that survives every estimator below. The dataset-clustered geometric mean on these four alone is $\mathbf{41.8}\times$ against the knob's $1.39\times$, where the original four gave $7.9\times$ against $1.48\times$.

That $41.8\times$ needs a caveat we would rather state than have read off us. Arm A certifies a radius of *exactly zero* on 4.0–81.5% of instances here — 81.5% on `spambase`, where the bound column reads $366.6\times$ — so the denominator is largely a floor, and the ratio measures how far arm A has collapsed at least as much as how far arm D reaches. Restricted to the instances where arm A certifies anything at all — computed the same way, as a ratio of seed means — the comparison gives $\mathbf{21.5}\times$. That restriction is itself lossy: on `spambase` only 2 of the 5 seeds retain any instance with a positive arm-A radius (and 4 of 5 on `parkinsons_tm`), so the restricted cell for the widest dataset rests on 37 of its 200 instances. This truncation is the opposite of the $r_{\max}$ ceiling the rest of the paper discusses: the ceiling caps radii (and a ratio of capped means has no fixed direction, App. A49), while the floor inflates these ratios and runs *for* us. Both estimators are far above the knob's $1.39\times$ and the ordering is the same, which is the claim; the magnitude is not stable to the choice and we do not rest anything on it. The pooled $18.2\times$ survives the same question (12.2–$19.5\times$ across estimators); the body quotes the per-cluster figures, not this one.

So the finding is not an artifact of the four datasets it was discovered on. These four are the harder test in the sense that the frontier was not developed on them — though three of them (`concrete`, `credit_default`, `spambase`) do appear elsewhere in this paper, and only `parkinsons_tm` is new to it entirely.

And the gap grows with d — because arm A collapses. Controlling for the arm-A radius, d has no residual association with the bound effect, so this is one finding and not two: arm D's radius *falls* over the same range (0.60 to 0.13), and it is the denominator that collapses. The bound effect runs 6.2, 50.0, 26.9, $366.6\times$ at $d = 8, 19, 23, 57$ — not monotone (`credit_default` sits below `parkinsons_tm`) but unmistakable in direction. The mechanism is the one the paper argues throughout: arm A discards the sign of every excluded term, so it throws away an amount that grows with the number of excluded terms, and at d=57 with $|S|/d$=0.69 there are still 18 of them. This is worth stating against our own scaling result (App. A40): as d grows the *explanation* degrades — $|S|/d$ rises from 0.42 to 0.69 here — while the *bound's* contribution grows. The certificate scales; the explanation does not; and the gap between the two bounds is widest exactly where the explanation is least useful. We do not read $366.6\times$ as a headline: it is a ratio against an arm-A radius of 0.0003, which is near-zero for the reason §5.3 gives, and a ratio to a near-zero denominator is unstable. The ordering is the claim.

A27 Coupling Diagnostic: Full Protocol

We turn that multimodal coupling pattern into an *exploratory* diagnostic — exploratory because the preregistered H1 criteria ($n \geq 30$, ≥ 3 domains) are not met. Train two matched-capacity surrogates, **JointMLP** on $[\mathbf{x}_{\text{tab}}; \mathbf{x}_{\text{mod}}]$ and **AdditiveMLP** $h_{\text{tab}} + h_{\text{mod}}$, and define $C(\mathcal{D}) := R^2_{\text{joint}} - R^2_{\text{additive}}$, which is 0 iff the DGP is additively decomposable across modalities. Across $n = 22$ benchmarks the realised gap $\Delta = R^2_{\text{E2E}} - R^2_{\text{Late}}$ correlates with C at Spearman $\rho = +0.766$ ($p < 10^{-4}$); but that pool mixes multimodal cells with tabular proxies, and on the 8 genuinely multimodal cells alone it is only $\rho = +0.500$ ($p = 0.21$), so we rely on the controlled experiment below.

Controlled amplitude sweep (preregistered, P3). On 7 (dataset, encoder) series — PathMNIST and DermaMNIST (Yang et al., 2023) and EuroSAT with Conv4 and with frozen DINOv2, plus 20-Newsgroups with frozen MiniLM — we inject a cross-modal coupling of tunable amplitude $\alpha \in \{0, 0.5, 1\}$ into the label and measure both C and the realized Δ per cell. Within *every* series, raising α raises *both* C and Δ monotonically (7/7 in each; mean within-series $\rho(C, \Delta) = 1.0$), and the $\alpha = 0$ additive controls sit at $C, \Delta \approx 0$ (mean $|C| = 0.006$, $|\Delta| = 0.001$) exactly as the diagnostic predicts. Pooled over the 14 coupled cells $\rho(C, \Delta) = +0.690$ ($p = 0.006$; domain-cluster bootstrap 95% CI $[0.25, 1.00]$; leave-one-domain-out $\rho = 0.62/0.76/0.74$). This is *controlled interventional* evidence (synthetic labels, real inputs) on multimodal data, not an observational proxy correlation, and the α knob is monotone in both C and Δ on 7/7 series. *Scope (honest)*: the pool above spans

three domains — medical-vision, remote-sensing, and text (20-Newsgroups + frozen MiniLM) — but our own registration excludes the text cells from the coupling headline (Deviations, 2026-07-12, recorded when the text probe came back negative), so on the registration's own terms this is a **two**-domain result and the $\geq$3-domain criterion is *not* met. Dropping those cells is the arm that flatters us — $\rho = 0.735$ against the 0.690 we quote — which is why the pooled correlation we report keeps them. Either way $n = 14$ coupled cells falls short of the preregistered $n \geq 30$, and the sign-prediction AUROC remains *undefined* because every coupled cell has $\Delta > 0$ (no negative class). H1 therefore remains **not met** and the result stays exploratory; the audio domain is deferred.

Held-out validation (I6). On an out-of-calibration split ($C^* = +0.005$ fixed on a 16-dataset calibration pool, $\rho = +0.797$, never touching the held-out sets; each held-out set predicted from its C alone), held-out sign accuracy on the 6 new tabular datasets is $4/6$ — modest and not individually significant (n=6, $p \approx 0.34$), so we lean on the full-pool ρ; both misses (`concrete`, `diabetes_reg`) have $|C| < 0.025$, inside the threshold noise band. Reproduction scripts and the legacy single-validation headline are in the supplement.

Theoretical anchor. App. A28.1 grounds C as the modality-block Hoeffding–Sobol cross-modal interaction index (Hoeffding, 1948; Sobol, 1993), which at the population level equals the normalised late-fusion gap of Theorem 4; unlike Friedman's H-statistic (Friedman & Popescu, 2008) or NID (Tsang et al., 2018), C requires only one additive + one joint MLP. *Practitioner guidance:* compute C once (~ 1 GPU-min/dataset); deploy end-to-end VIB-NAM when $C > 0$ (LOOCV-optimal); a cross-FM check (DINOv2/SigLIP) shows DermaMNIST is a consistent $C < 0/\Delta > 0$ outlier — C is necessary-not-sufficient cross-FM.

A28 The Representability Gap, Stated Exactly

The body defers this statement here (§A3). An earlier revision stated it as a quantitative lower bound in terms of the cross-modal Jacobian; that bound is false, and the counterexample is given at the end of this appendix. What holds is an identity.

Setup. Consider a multimodal supervised problem on $\mathbb{X}_{\text{tab}} \times \mathbb{X}_{\text{mod}} \to \mathbb{R}$ with population distribution $\mathcal{P}$ under which $X_{\text{tab}} \perp X_{\text{mod}}$, and Bayes-optimal regressor $\hat{y} \in L^2(\mathcal{P})$. A *late-fusion predictor* has the form $\hat{f}_{\text{LF}}(x_{\text{tab}}, x_{\text{mod}}) = \Phi(g_{\text{tab}}(x_{\text{tab}}),\, g_{\text{mod}}(x_{\text{mod}}))$ with single-modality predictors $g_{\text{tab}}, g_{\text{mod}}$ and the linear combiner $\Phi(u, v) = \alpha u + \mu v + \gamma$ (the mixing weight μ is distinct from the IB compression β_i). Write the Hoeffding decomposition (Hoeffding, 1948) $\hat{y} = \hat{y}_\emptyset + \hat{y}_{\text{tab}} + \hat{y}_{\text{mod}} + \hat{y}_{\text{tab,mod}}$, whose components are mean-zero and mutually L^2-orthogonal.

Lemma 1 (Late fusion is additive across modalities)**.** *Every late-fusion predictor with a linear combiner is of the form $a(x_{\text{tab}}) + b(x_{\text{mod}})$; where it is twice differentiable, $\partial^2 \hat{f}_{\text{LF}} / (\partial x_{\text{tab},i}\, \partial x_{\text{mod},j}) = 0$ for all i, j.*

Proof. Φ is affine, so $\hat{f}_{\text{LF}} = \alpha g_{\text{tab}} + (\mu g_{\text{mod}} + \gamma)$. □ □

Theorem 4 (Representability gap)**.** *Under the setup above,*

$$\inf_{\hat{f}_{\text{LF}}} \mathbb{E}\big[(\hat{y}(X) - \hat{f}_{\text{LF}}(X))^2\big] \;=\; \text{Var}\big(\hat{y}_{\text{tab,mod}}(X)\big),$$

the infimum over all late-fusion predictors with a linear combiner. If $\mathcal{P}$ has a continuous positive density on a compact product of intervals and $\hat{y}$ is C^2 there, the gap is positive if and only if the cross-modal mixed derivative $\partial^2 \hat{y} / (\partial x_{\text{tab}}\, \partial x_{\text{mod}})$ is not identically zero. A joint head that is a universal approximator can attain $\hat{f} = \hat{y}$, so for it the gap vanishes. VIB-NAM's pairwise interaction phase (§A3) closes the gap only when $\hat{y}_{\text{tab,mod}}$ is itself a sum of pairwise terms: a sum of pairwise blocks cannot represent, e.g., $x_1 x_2 x_3$, which under independent mean-zero inputs is orthogonal to every main and pairwise effect.

Proof. By Lemma 1 every $\hat{f}_{\text{LF}}$ lies in the closed subspace $\mathcal{A} = \{a(X_{\text{tab}}) + b(X_{\text{mod}})\}$, and every element of $\mathcal{A}$ is a late-fusion predictor ($g_{\text{tab}} = a$, $g_{\text{mod}} = b$, $\alpha = \mu = 1$, $\gamma = 0$), so the infimum is the squared distance from $\hat{y}$ to $\mathcal{A}$. Under independence the orthogonal projection onto $\mathcal{A}$ is $\hat{y}_\emptyset + \hat{y}_{\text{tab}} + \hat{y}_{\text{mod}}$ (Hoeffding, 1948), leaving the residual $\hat{y}_{\text{tab,mod}}$, which is mean-zero; its second

moment is its variance. For the characterisation: if the mixed derivative vanishes identically on the (connected) product domain, integrating it gives $\hat{y} = a + b$ there, so $\hat{y}_{\text{tab,mod}} = 0$. Conversely, if $\hat{y}_{\text{tab,mod}} = 0$ in L^2, then $\hat{y} = a + b$ almost everywhere with a, b the (continuous, by continuity of $\hat{y}$ and of the density on a compact domain) conditional-mean components, hence everywhere on the support by continuity, and the mixed derivative is zero. □ □

The coupling diagnostic $C(\mathcal{D})$ of App. A27 estimates exactly this gap in normalised form: Prop. 2 below shows $C^\star = \text{Var}(\hat{y}_{\text{tab,mod}})/\text{Var}(Y)$ when the fitted heads reach their population optima ($\rho = +0.766$, $n = 22$, against the observed end-to-end advantage).

Nonlinear combiners. If Φ may be nonlinear, e.g. $\Phi(u, v) = \alpha u + \mu v + \gamma uv$, the late-fusion predictor is no longer additive and the theorem does not apply. The multimodal benchmarks this theorem was written for (cut from this revision for length; runners and artifacts ship with the code) used a linear Φ (OLS or 1-D logistic).

The withdrawn bound, and why it is false. The earlier revision claimed $\mathbb{E}[(\hat{y} - \hat{f}_{\text{LF}})^2] \geq J(\hat{y})/(8L^4)$ with $J(\hat{y}) := \mathbb{E}\|\partial^2 \hat{y}/\partial x_{\text{tab}} \partial x_{\text{mod}}\|_F^2$ and L a domain bound. Take independent $X, Y \sim \text{Unif}[-1, 1]$ ($L = 1$), $k = n\pi$ and $f_n(x, y) = \sin(kx)\sin(ky)/(2k^2)$. Both conditional means vanish, so the best additive approximation is 0 and the gap is $\mathbb{E}[f_n^2] = 1/(16k^4) \to 0$, while $J(f_n) = 1/16$ for every n and the Hessian operator norm stays $\leq 1/2$. The claimed right side, $1/128$, exceeds the true gap by $12.2\times$ already at $n{=}1$. No bound of the form gap $\geq c\,J$ with $c > 0$ can hold under a Hessian bound alone: a mixed derivative controls the gap only together with a frequency (higher-derivative) bound, because high-frequency interactions carry large derivatives and small variance. The Sobolev-embedding step of the old proof ran the inequality in the wrong direction, and its Taylor step assumed a point at which the late-fusion fit matches $\hat{y}$ in value and gradient, which least squares does not supply. We withdraw the bound, the "tight constant" paragraph, and the figure that compared an empirical gap to it: a constant ratio to a false bound is not evidence about the bound.

A28.1 Sobol Interpretation of $C(\mathcal{D})$

Hoeffding–Sobol decomposition. Let $Y = f(X_{\text{tab}}, X_{\text{mod}}) + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$, $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$, and $\text{Var}(Y) < \infty$, with $X_{\text{tab}} \perp X_{\text{mod}}$ (assumed throughout this subsection; correlated-modality extensions follow generalised Sobol (1993) indices). The Hoeffding ANOVA decomposition (Hoeffding, 1948) writes

$$f(x_{\text{tab}}, x_{\text{mod}}) = f_\emptyset + f_{\text{tab}}(x_{\text{tab}}) + f_{\text{mod}}(x_{\text{mod}}) + f_{\text{tab,mod}}(x_{\text{tab}}, x_{\text{mod}}),$$

with $\mathbb{E}[f_S(X_S)] = 0$ for every non-empty subset $S \in \{\text{tab}, \text{mod}, (\text{tab}, \text{mod})\}$ and L^2-orthogonality $\mathbb{E}[f_S\, f_{S'}] = 0$ for $S \neq S'$. The *Sobol cross-modal interaction index* is

$$S_{\text{tab,mod}} := \frac{\text{Var}(f_{\text{tab,mod}}(X_{\text{tab}}, X_{\text{mod}}))}{\text{Var}(Y)} \in [0, 1].$$

Proposition 2 (Coupling as Sobol cross-modal index)**.** *Let $h^\star_{\text{joint}}$ and $h^\star_{\text{add}} = h^\star_{\text{tab}} + h^\star_{\text{mod}}$ denote the population-optimal joint and best additive (across-modality) predictors under squared loss. Then*

$$C^\star := R^2_{\text{joint}}(\mathcal{D}) - R^2_{\text{additive}}(\mathcal{D}) \xrightarrow{p} S_{\text{tab,mod}}$$

as the fitted heads approach their population optima and $|\mathcal{D}| \to \infty$. Hence $C^\star \geq 0$, with equality iff the representability gap of Theorem 4 is zero.

Proof sketch. By orthogonality, the population L^2-projection of f onto the additive subspace $\{a(X_{\text{tab}}) + b(X_{\text{mod}})\}$ is exactly $f_\emptyset + f_{\text{tab}} + f_{\text{mod}}$, so $h^\star_{\text{add}}(X) = f_\emptyset + f_{\text{tab}} + f_{\text{mod}}$ and $h^\star_{\text{joint}}(X) = f$, both Bayes-optimal in their class. The residual variances are $\mathbb{E}[(Y - h^\star_{\text{joint}})^2] = \sigma_\varepsilon^2$ and, by orthogonality, $\mathbb{E}[(Y - h^\star_{\text{add}})^2] = \text{Var}(f_{\text{tab,mod}}) + \sigma_\varepsilon^2$. Substituting into $R^2 = 1 - \mathbb{E}[(Y - \hat{h})^2]/\text{Var}(Y)$,

$$C^\star = \frac{\text{Var}(f_{\text{tab,mod}})}{\text{Var}(Y)} = S_{\text{tab,mod}}.$$

No further rescaling enters: both R^2 values share the denominator $\text{Var}(Y)$, and σ_ε^2 cancels in the difference. Finite-capacity heads that fall short of their optima add an estimation error of either sign, which is why C is used only through rank correlations. □ □

Discussion. Proposition 2 positions $C(\mathcal{D})$ as the modality-block instantiation of Sobol–Hoeffding total interaction: at the population level it equals the normalised late-fusion gap, with no dataset-dependent constant; rank correlations across datasets (the only use we make of C in App. A27) are computed on this scale. The practical innovation is the two-MLP estimator – one additive, one joint – which sidesteps the intractable joint partial-dependence integrals required by Friedman & Popescu (2008) when X_{mod} is a D-dim FM embedding ($D \sim 512$–1024).

A29 Withdrawn β–MI Bound, and What Replaces It

Retraction. We do not claim the β–conditional-MI lower bound one might hope to state here, of the form $I(X_i; Y \mid X_{-i})/(s_i + \epsilon) \geq \Psi(\beta_i)$ with $\Psi(\beta_i) = \log(1 + e^{-\tilde{\beta}_i})$, and we had conjectured it. It fails in three independent ways, recorded so the error is not repeated. (i) Its first step applied the data-processing inequality to $X_i \to Z_i \to f_i(X_i)$, but $f_i(x_i) = \text{Dec}(z_i) + w_i x_i + b_i$ reads x_i directly, so this is not a Markov chain; for $w_i \neq 0$ the map is generically injective and $I(X_i; f_i(X_i)) = H(X_i)$, independent of β_i. (ii) Its KKT lemma dropped the term $\beta_i \, \partial I_q / \partial \tilde{\beta}_i$, which the envelope theorem says cancels against $\partial \mathcal{L}_{\text{pred}} / \partial \tilde{\beta}_i$ at a θ-optimum; carrying the paper's own substitution through gives $I_q = s_i$, with no $1/\text{sigmoid}(\tilde{\beta}_i)$ factor — and that factor was the sole source of β-dependence in the conclusion. (iii) The resulting inequality is vacuous as $\beta_i \to \infty$ ($\Psi \to 0$, so it reduces to $I \geq 0$) and false in the regime we actually operate in: at the reported $\bar{\beta} \in [0.020, 0.035]$ one has $\Psi \in [3.4, 3.9]$ nats, so it would assert $I \geq 3.4\, s_i$, which the three planted pure-noise features of our own d=10 synthetic design ($I = 0$, $s_i = 0$) violate.

The frontier is unchanged. The correct scoped statement is the following, and it is the only one we now rely on.

Proposition 3 (The bottleneck constrains the residual channel)**.** *Write $f_i(x_i) = g_i(x_i) + w_i x_i + b_i$ with $g_i(x_i) = \text{Dec}(z_i)$ the stochastic path. Then $X_i \to Z_i \to g_i(X_i)$ is a Markov chain, so by data processing $I(X_i; g_i(X_i)) \leq I(X_i; Z_i) \leq R_i$, where $R_i = \mathbb{E}_x \text{KL}(q(z \mid x) \| p(z))$ is the variational rate. Consequently $R_i \to 0$ forces the residual channel to carry no information about X_i and f_i collapses to the affine bypass (Thm. 1), while placing* no *constraint on $I(X_i; f_i(X_i))$, which the bypass keeps maximal.*

This is what licenses reading R_i as a *nonlinearity* diagnostic (Cor. 1) rather than an importance or information-contribution diagnostic. It says nothing about the total information f_i carries about X_i, which the bypass alone already makes maximal; the scope is the residual channel and nothing wider.

A30 High-Dimensional VIB-X Demo

To probe VIB-X value on larger feature spaces, we run it on three datasets with $d \in \{16, 23, 30\}$ — wider than the frontier's $d \leq 13$, though narrower than App. A26's d=57, so "high-dimensional" here means only "wider than the frontier". The two OpenML sets are 44065 (*nyc-taxi-green-dec-2016*; the artifacts carry it under the legacy key `superconduct`) and 42477 (*default-of-credit-card-clients*). We sweep $\varepsilon \in \{0.05, 0.1, 0.2, 0.5\}$. Table 20 reports the mean number of features retained and the dimensionality reduction $1 - \bar{k}/d$. On `nyc_taxi` and `credit_default`, VIB-X at $\varepsilon = 0.5$ cuts feature reporting by 57–62%; on breast_cancer (where the classifier truly uses most features) the reduction is modest (30% at $\varepsilon = 0.5$), which is honest feedback that VIB-X is not a magic sparsifier when the underlying model is genuinely dense. This table is at the paper's G=10^6 grid; built at G=500 knots instead, every cell is unchanged to two decimals, because the fit error of the simple forms dominates the $L_i h/2$ padding on these datasets. The qualitative examples of App. A32 are still the G=500 artifact, where the larger padding makes simple forms harder to afford, so VIB-X needs *more* features than it would on a dense table; we checked that the choice set did not degenerate there — constant and linear forms are selected throughout, unlike the latency run this grid did spoil (App. A14) — and left those examples as they are.

Table 20: Mean number of features retained by VIB-X greedy at varying tolerance on high-dimensional datasets ($n = 100$ test samples each, `results/atlas_highd/atlas_highd_seed42.json`). Reduction = $1 - \bar{k}/d$.

Dataset	d	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.5$
`nyc_taxi`	16	13.89 (13%)	12.78 (20%)	10.81 (32%)	6.12 (62%)
`credit_default`	23	18.70 (19%)	16.98 (26%)	14.39 (37%)	9.94 (57%)
breast_cancer	30	27.46 (8%)	26.23 (13%)	24.54 (18%)	21.05 (30%)

A31 SHAP VS VIB-X LATENCY

We directly compare per-sample wall-clock inference latency of `shap.KernelExplainer` against VIB-X greedy on the four datasets of App A13. Both are measured on the same trained VIB-NAM with a background set of 50 samples for SHAP. VIB-X is **56–214**× faster: california (d=8) SHAP 15.8 ms vs VIB-X 0.28 ms (56×); concrete (d=8) 16.2 ms vs 0.27 ms (59×); diabetes_reg (d=10) 48.5 ms vs 0.33 ms (145×); heart (d=13) 113.6 ms vs 0.53 ms (214×) (`results/atlas_latency/shap_vs_atlas_seed42.json`). A ratio between two wall-clock measurements is only meaningful within one environment — a different machine and `shap` release gave 170–420× for the same code, with `KernelExplainer` four times slower and VIB-X unchanged — which is why every artifact records its library versions (`results/ENVIRONMENT.json`, `elasticnam/utils_provenance.py`).

Deletion-test AOPC: what is and is not comparable. We report mean AOPC across 4 classification tasks (5 seeds × 15 explained samples) as VIB-X 4.19, LIME 3.86, SHAP 0.30 (Table 21), but only the VIB-X/LIME contrast is a like-for-like comparison: both rank features for VIB-NAM and are scored by masking *through VIB-NAM*. The SHAP arm is TreeSHAP on a separately trained XGBoost, scored by masking through *that* model, so its AOPC is expressed in a different predictor's sensitivity scale and its magnitude is *not* commensurable with the other two — the 4.19 vs. 0.30 gap should not be read as a faithfulness win. We report it only as a cross-model reference point. The same-model comparison that is meaningful, VIB-X vs. LIME, is a $+8.4\%$ AOPC margin on the grand mean ($+4.0\%$ to $+10.2\%$ across the four datasets). We also note that for an *additive* model the exact Shapley value of feature i is $f_i(x_i) - \mathbb{E}[f_i(X_i)]$ — for the *interventional* (marginal) value function with the training distribution as background, the one KernelSHAP estimates; the *conditional* value function does not have this closed form under dependent features. VIB-X ranks by the raw $|f_i(x_i)|$, which is not this value unless the shapes are centred, so the two rankings differ by a gauge choice that App. A50 measures. Both are computable in $O(d)$; the 56–214× latency figures above are therefore against `KernelExplainer`, not against that closed form, and should be read as a statement about sampling-based estimators only.

Table 21: VIB-X greedy k-feature explanations vs. SHAP and LIME top-k on deletion-test AOPC (5 seeds × 15 samples, $\varepsilon_{\text{frac}} = 0.1$). Higher = explanation identifies more prediction-critical features. **Only the VIB-X/LIME columns are commensurable** (both rank features for VIB-NAM and are scored by masking through VIB-NAM); the SHAP column is TreeSHAP on a separately trained XGBoost scored through *that* model and is reported as a cross-model reference only — its magnitude must not be read as a faithfulness comparison (see text).

Dataset	VIB-X	SHAP	LIME
heart	$\mathbf{1.89 \pm 0.43}$	0.35 ± 0.03	1.72 ± 0.39
breast_cancer	$\mathbf{11.33 \pm 4.78}$	0.34 ± 0.05	10.41 ± 4.41
adult	$\mathbf{2.20 \pm 0.31}$	0.20 ± 0.03	2.06 ± 0.33
diabetes_class	$\mathbf{1.32 \pm 0.32}$	0.31 ± 0.03	1.27 ± 0.31
mean	**4.19**	0.30	3.86

A32 VIB-X Qualitative Examples

To make VIB-X output concrete, Table 22 shows three explanations per dataset (california, concrete, diabetes_reg) at tolerance $\varepsilon = 0.2$. For each test point, we list the selected features, their chosen complexity level (constant / linear / quadratic / full), and per-feature contribution $f_i(x_i)$. The full JSON is in `results/atlas_qualitative/atlas_qualitative_seed42.json`.

Table 22: VIB-X explanations at $\varepsilon = 0.2$. "used" is the number of features retained in the explanation out of d total; the level mix is a concise fingerprint (F=full, Q=quadratic, L=linear, C=constant).

Dataset	d	Sample	pred	target	used/d	level mix
california	8	2702	+1.19	4.64	6/8	5F + 1C
		368	+0.55	2.18	5/8	4F + 1L
		3194	−0.49	1.65	6/8	5F + 1Q
concrete	8	134	−1.86	8.20	4/8	4F
		18	+0.20	37.17	6/8	5F + 1L
		158	+0.70	55.90	5/8	4F + 1Q
diabetes_reg	10	58	−0.54	168.0	6/10	5F + 1Q
		7	+0.27	272.0	6/10	5F + 1Q
		68	+0.07	140.0	6/10	5F + 1Q

VIB-X regularly halves the number of reported features (e.g., $4/8$ on concrete, $5/8$ on california) while preserving approximation error below ε; 1–2 features receive simplified (linear or quadratic) approximations on most predictions.

A33 Backbone-Strength Sweep Figure

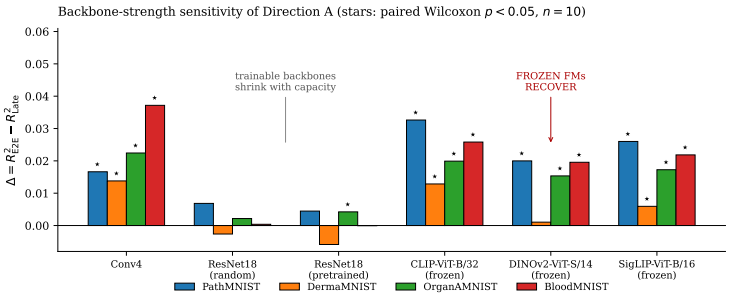

Figure 9: **Backbone-strength sensitivity (Stream A1).** $\Delta = R^2_{\text{E2E}} - R^2_{\text{Late}}$ across 4 image encoders $\times$ 4 medmnist datasets ($n = 10$, stars: paired Wilcoxon $p < 0.05$). Trainable backbones: monotone narrowing in encoder capacity. Frozen FM backbone (CLIP): gap recovers, consistent with an FM-augmented-tabular regime where end-to-end glass-box composition is required.

A34 Architecture Diagram

A35 VIB-X Optimality Table

A36 The VIB-X Program, Its Relaxation, and the Greedy

Everything VIB-X optimises, in one place. Fix x and the gauge in which the shapes are read (raw by default; App. A50). Let $a_i = |f_i(x_i)|$ and let e_i^l be the error-table entry of feature i at level $l \in \{0, 1, 2, 3\}$ = {constant, linear, quadratic, `full`} ($e_i^3 = 0$ up to padding).

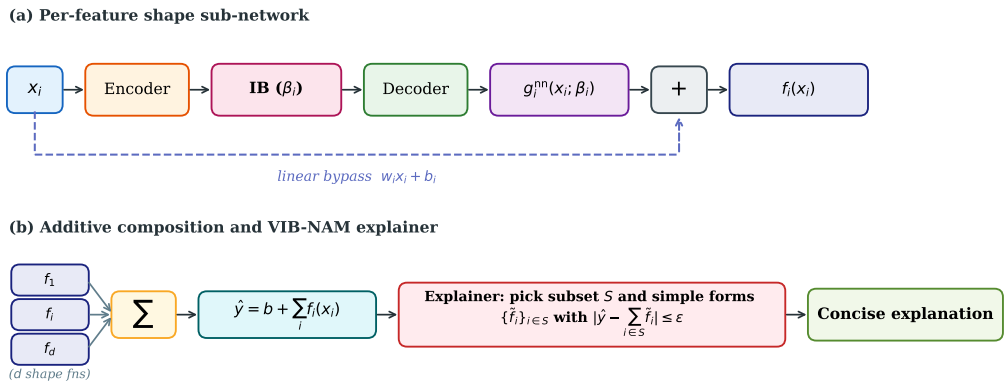

Figure 10: Overview of VIB-NAM architecture. Each feature x_i passes through an encoder–bottleneck–decoder path (controlled by β_i) plus a linear bypass $w_i x_i + b_i$. Shape function outputs are summed additively. When β_i is large, the bottleneck compresses the nonlinear path, and the model gracefully degrades to a linear function of x_i via the bypass (Theorem 1).

Table 23: VIB-X greedy vs. exact exhaustive search / MIP, computed on 100 randomly sampled test predictions per dataset (heart n=54, diabetes_reg n=89) with a fixed selection seed (artefacts: `results/atlas_mip/atlas_mip_eps0.1_seed42.json`, `results/atlas_analysis/atlas_optimality_seed42.json`). Ratio = greedy cost / exact cost ($\geq$ 1.0; 1.0 = optimal). % Exact = fraction of predictions where greedy matches the verified-Optimal exact solution (only PuLP/CBC "Optimal" status is counted as the exact reference). Mean MIP solve time is $\sim$ 0.02 s / sample. Both arms use a G=2×10^5 error table — the grid the MIP study was run at, coarser than the 10^6 the certificates elsewhere use. App. A14 records what a coarser table does to this comparison — at G=500 the padding $L_i h/2$ exceeds ε for every feature, every simple form is infeasible, and both solvers are timed on a degenerate program — and the gate exempts these two artifacts by name for that reason.

Dataset	d	Mean Ratio	Max Ratio	% Exact	Reference
california	8	1.000	1.000	100.0%	exhaustive
concrete	8	1.009	1.040	72.0%	exhaustive
abalone	8	1.000	1.000	100.0%	MIP
diabetes_reg	10	1.005	1.030	64.0%	MIP
wine	11	1.000	1.000	100.0%	MIP
bike	12	1.009	1.027	51.0%	MIP
heart (cls)	13	1.000	1.014	98.1%	MIP

The integer program. Binary z_i (feature i shown) and $c_{i,l}$ (shown at level l):

$$\min \; w_f \sum_i z_i + w_c \sum_{i,l} l\, c_{i,l} \quad \text{s.t.} \quad \sum_l c_{i,l} = z_i \;\forall i, \qquad \sum_i (1 - z_i)\, a_i + \sum_{i,l} c_{i,l}\, e_i^l \le \varepsilon,$$

with w_f=1 and w_c=0.1 throughout (`elasticnam/atlas/mip_solver.py`). The objective is a *cost*, not a feature count: a feature at `full` costs 1.3, at constant 1.0. The constraint is the conservative surrogate of §3.2; the signed variant (arm C) replaces $\sum_i (1 - z_i) a_i$ by $|\sum_i (1 - z_i) f_i(x_i)|$, which is still a MILP: with $u = \sum_i (1 - z_i) f_i(x_i)$ and $E = \sum_{i,l} c_{i,l} e_i^l$, the constraint $|u| + E \le \varepsilon$ is the pair $u + E \le \varepsilon$, $-u + E \le \varepsilon$. We did not solve it; arm C uses the greedy.

Its structure. Writing the constraint as a covering requirement — choose options whose reductions $r_{i,l} = a_i - e_i^l$ sum to at least $R = \sum_i a_i - \varepsilon$, at most one option per feature — it is a multiple-choice covering knapsack. The body's earlier description as weighted set cover was loose: an element is not covered by a set, a scalar deficit is covered by additive reductions, and the $(1 + \ln d)$ set-cover guarantee does not transfer to the greedy we run.

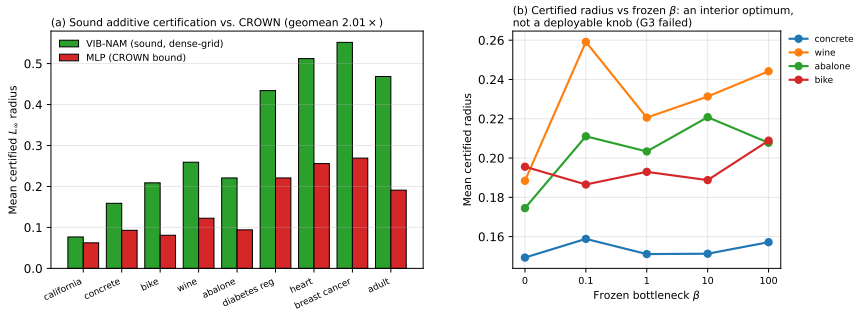

Figure 11: **Sound additive certification (§4).** (a) Mean certified L_∞ radius: VIB-NAM (sound dense-grid interval with certified analytic-Lipschitz padding; *best-β oracle envelope* shown — geometric mean $2.01\times$ *on seed* 42, the seed this figure is drawn from; the five-seed validation-selected headline is $1.87\times$, §4) vs. the *lower bound* CROWN can prove for the regularized MLP baseline (not parameter-matched; see text). (b) Mean certified radius vs. the bottleneck β on regression datasets: a moderate β lifts the radius (the certified-radius *lower bound* is empirically monotone in compression; the supporting argument in App. A22 is heuristic, not a theorem), with an interior optimum at $\beta \in \{1, 10\}$ rather than monotone growth.

The relaxation and the certificate. Relaxing $c_{i,l} \in [0, 1]$ gives a fractional multiple-choice knapsack, solved exactly by taking, per feature, the lower convex hull of its (reduction, cost) options and consuming hull segments in increasing cost-per-reduction order until R is met (`elasticnam/atlas/lp_bound.py`, $O(dL \log dL)$). Its value LP satisfies LP $\leq$ OPT $\leq$ cost(greedy), so cost/LP upper-bounds the greedy's ratio to the optimum of *this* program. The "within 0.9% of the MIP optimum" of §5.2 is a ratio of these costs (1.009 mean), measured against CBC's solution, not a ratio of feature counts; the per-instance LP certificate is the looser, solver-free version of the same statement.

The greedy.

1. Sort features by a_i, descending. Starting from $S = \emptyset$, while $\sum_{i \notin S} a_i > \varepsilon$: add the next feature at the lowest level l for which the constraint holds, else `full`.
2. For each $i \in S$, lower its level while the constraint holds.
3. For each $i \in S$, drop it if the constraint still holds without it.

The ranking in step 1 reads $|f_i(x_i)|$ in whatever gauge the shapes carry; it is not the Shapley ranking unless the shapes are centred (App. A31).

A37 Main Benchmark Table

What VIB-NAM-Boost-Ens is. The second-ranked row of Table 24 was named and never defined, which an audit caught. Round 0 is a full VIB-NAM; rounds $1..T$ each add a per-feature tiny MLP (32 hidden, one layer) fitted to the running residual under shrinkage η=0.1, and pairwise interactions are discovered on the final residual:

$$\hat{y} = VIB - NAM(x) + \sum_{t=1}^{T} \eta \sum_{i} \delta_{t,i}(x_i) + \sum_{i<j} h_{ij}(x_i, x_j).$$

"Ens" averages over seeds. Implementation: `elasticnam/models/elastic_nam_boost.py`; its η, T and round width are swept in App. A12. The model remains a sum of one-dimensional terms plus lookup-table pairs, so the certificates apply to it as to base VIB-NAM: each 2-D block is exact (Prop. 1), but because the pairs share coordinates with the main effects the model-level interval and arm D are sound, not attained (Prop. 4 needs a main-effects model).

Table 24: Main benchmark under nested CV (outer 5-fold; inner 20% validation for selection; held-out outer fold reported). R^2 for regression, AUROC for classification. **Best bold**, second underlined. VIB-NAM-B = VIB-NAM-Boost, VIB-NAM-B-Ens = 3-member ensemble; NAM is a vanilla per-feature MLP (no IB, no bypass). Avg. Rank is over all eight methods on nine datasets. **Bold marks the largest point estimate and is not a significance claim**: per-fold SDs reach ± 0.0865, and on 4 of the 9 datasets the winner is inside one SD of the runner-up (`heart` is 0.9044 against 0.9036, a gap of 0.0008 against an SD of 0.032). §5.1 reports what survives a paired test. Cells and the rank row are regenerated by the table gate.

Dataset (task)	VIB-NAM	VIB-NAM-B	VIB-NAM-B-Ens	NAM	NODE-GAM	EBM	XGBoost	Linear
california (reg)	0.794	0.802	0.803	0.748	0.800	0.836	**0.847**	0.631
heart (cls)	0.904	0.903	0.904	**0.904**	0.894	0.886	0.857	0.901
breast_cancer (cls)	0.994	0.993	0.994	0.995	**0.997**	0.995	0.994	0.995
diabetes_reg (reg)	0.475	0.468	0.469	0.483	0.373	**0.489**	0.357	0.472
concrete (reg)	0.910	0.919	0.921	0.857	0.905	**0.933**	0.922	0.602
bike (reg)	0.841	0.894	0.895	0.693	0.897	0.923	**0.952**	0.394
wine (reg)	0.421	0.431	0.434	0.330	0.377	0.415	**0.473**	0.290
abalone (reg)	0.555	0.564	**0.566**	0.538	0.560	0.553	0.517	0.526
adult (cls)	0.919	0.919	0.919	0.916	0.915	**0.930**	0.927	0.853
Avg. Rank	4.39	4.56	3.28	5.22	4.89	**2.89**	4.22	6.56

A38 Positioning Against Formal XAI: The Full Comparison

Abductive / prime-implicant explanations (Ignatiev et al., 2019) fix the explanation's features at their instance values and guarantee the prediction for *any* value of the remaining features; La Malfa et al. (2021) and Wu et al. (2023) bound that perturbation, and Izza et al. (2024a) restrict the perturbed point to an L_p ball of radius ε around the instance — a sound per-instance radius on which an explanation remains a valid sufficient reason. The claim "no prior work certifies an explanation under bounded input perturbation" would therefore be false, and we do not make it. Nor is the choice binary between a sampled sensitivity rate and a hard certificate: probabilistic AXps and δ-relevant sets (Wäldchen et al., 2021; Izza et al., 2023) give a formal *quantitative* guarantee — sufficiency on a stated measure of completions. Our frontier trades $|S|$ against a guarantee level and is structurally the same kind of object under a different measure (a ball radius rather than a completion measure), which is worth saying explicitly rather than leaving the reader to notice. Our statement differs on three axes.

The two regression guarantees, with their quantifiers. Bassan et al. (2026) treat regression explicitly (their App. E.2, Algs. 6–8). Their subset S is sufficient for x at radius ϵ_p and deviation δ when

$$\forall \tilde{x} \in \mathcal{B}_p^{\epsilon_p}(x): \quad \left| f(x_S; \tilde{x}_{\bar{S}}) - f(x) \right| < \delta,$$

a statement about the *model* with x_S held at its instance value. Ours, for a displayed surrogate $g_{S,\mathcal{C}}(x) = b + \sum_{i \in S} g_i^{\mathcal{C}_i}(x_i)$, is

$$\forall x' \in \mathcal{B}_\infty^r(x): \quad |f(x') - g_{S,\mathcal{C}}(x')| \le \varepsilon,$$

with every coordinate free, including those in S, and the reference a simplified function evaluated at the perturbed point rather than the fixed number $f(x)$. They coincide only under two conditions together: forms at `full` ($g_i = f_i$ on S) *and* the surrogate's intercept anchored at the instance, $b + \sum_{i \notin S} f_i(x_i)$. Then ours reads $|\sum_{i \notin S}[f_i(x'_i) - f_i(x_i)]| \le \varepsilon$ over the ball, which is theirs with $\delta = \varepsilon$ (the coordinates in S drop out). With the plain intercept b they differ even at `full`: for $f = x_1 + x_2$, $x = (0, 1)$, $S = \{1\}$ and $r = 0.1$, the model moves by at most 0.1 while $g(x') = x'_1$ misses f by up to 1.1. So regression is not a distinction; the object is: we certify how well a *simpler function* tracks the model across a ball, they certify how little the model moves when the unnamed features do.

The two predicates, measured on the same explanations. We certified both on the headline's models and minimal explanations (`elasticnam/experiments/run_anchored_control.py`; `results/explanation_stability/anchored_control_seed42.json` and seeds 43–46): ours with the fixed intercept (arm D); the instance-anchored surrogate with the same forms; the anchored surrogate with every named form at `full` — their predicate, on our S; and the

anchored *empty* explanation, which displays only $f(x)$ and so certifies nothing but that the prediction does not move by ε.

Table 25: Share of minimal explanations certified at $r \geq 0.1$ / $r \geq 0.25$ (%), five seeds, ε=0.5, arm D throughout. The "fixed" column reproduces the headline's arm-D radii exactly.

Dataset	fixed intercept	anchored	anchored, `full`	anchored, empty
`abalone`	94 / 72	100 / 90	100 / 90	100 / 5
`wine`	94 / 57	100 / 79	100 / 80	99 / 28
`diabetes_reg`	99 / 90	100 / 100	100 / 100	100 / 98
`heart`	92 / 70	100 / 100	100 / 100	87 / 0

Two things follow. Under anchoring, simplifying the forms is nearly free: the `full`-form column tracks the anchored one to within 0.007 in mean radius. And at our ε the anchored predicate barely discriminates: the empty explanation already meets it at $r \geq 0.1$ on 87–100% of instances, and at $r \geq 0.25$ on `diabetes_reg` on 98%. A deviation-from-$f(x)$ predicate at this tolerance is close to the prediction's own stability; the fixed-intercept surrogate must carry the prediction's *level*, not only its change, which is why it certifies less. That difference — a displayed function that approximates f rather than a set that leaves f unmoved — is what this paper's certificate adds (`scripts/anchored_control_summary.py`).

(i) What moves. The abductive line pins the explanation's own features at their instance values and perturbs only the omitted ones. "We perturb every coordinate, including those in S" is not by itself a distinction any more. *Inflated* explanations (Izza et al., 2024b) generalise an AXp so that each feature *in* S ranges over an interval rather than being pinned, and distance-restricted AXps (Izza et al., 2024a) already confine the perturbed point to an L_p ball. The honest narrower statement: that line inflates S to per-feature intervals chosen by solver search and certifies a *decision*, whereas we perturb S over a fixed L_∞ ball in closed form and certify a real-valued *functional form* to within ε. Pinning S is the right model of "the reasons I gave you still hold"; perturbing all of S is the right model of "the input I measured was slightly wrong". **Where axis (i) bites, precisely.** At the `full` level $f_i - g_i \equiv 0$ on S, so arm D's criterion does not depend on whether S moves at all, and our radius *is* a distance-restricted sufficiency radius — with an ε-approximation predicate in place of decision invariance, computed by the same interval query the abductive oracle uses. Axis (i) therefore has content only when a simplified form is displayed ($e_i^{\mathcal{C}} \neq 0$), which is the case the frontier is about and the case axis (ii) names; readers should take (ii) and (iii) as the distinction and (i) as a consequence of certifying forms. **And we should say what that is worth.** Holding S fixed and restoring every form to `full` — by the argument just given, exactly the move that turns our radius into a distance-restricted sufficiency radius — changes it by 1.02–$1.12\times$ under arm D (App. A43). So certifying forms rather than subsets is worth at most 12% of radius on these models, and what separates the two lines is the predicate and the cost, not the magnitude.

(ii) What is certified. That line certifies invariance of the predicted *class*. We certify that a *displayed surrogate* — a subset *and* a per-feature functional form (constant/linear/quadratic/full) — stays within ε of the model output. That is the object a stakeholder is actually shown, and it is what the stability–simplicity frontier trades against: because a functional form is part of what we certify, there is a simplicity knob to trade, which does not arise when the certified object is a subset.

(iii) What it costs. Those guarantees come from SAT/MILP search over general classifiers, and the cost is solver-dependent. That cost is the axis the formal line is itself attacking — De Palma et al. (2026) reuse verifier queries across features to push verified explanations to networks with 10^5 activations — and Bassan et al. (2025) show the complexity of the underlying question is not uniform across GAMs: it moves with the input domain, the component model class, and regression versus classification, and expressing a network additively helps only under some of those settings. For an additive model our radius is closed-form from per-feature Lipschitz structure and costs milliseconds, which is what makes the frontier sweep of App. A54 affordable at all. The distinction we draw is not that a solver is slow but that we never call one; the price is the LP bound in the optimality row below.

The nearest prior work, compared on stated axes. Bassan et al. (2026) target exactly our model class. The comparison below is on the two papers' stated guarantees and complexities, which

are directly comparable; we did *not* run their code, and say so rather than implying an empirical head-to-head.

Table 26: Our guarantee against Bassan et al. (2026), on the two papers' *stated* guarantees and complexities. We did not run their code and offer no empirical head-to-head.

	Bassan et al. (2026)	this paper
object certified	feature subset	subset *and* per-feature form
guarantee	cardinally *minimal*	*valid under* $\|\delta\|_\infty \le r$
what may move	omitted features	*every* coordinate
optimality	exact	LP bound on a surrogate
cost	$O(\log d)$ verification queries + preprocessing	$O(dL \log dL)$ closed form, no solver

Read across the rows, the two are complementary rather than competing: they answer "is this the smallest sufficient subset?" exactly, and we answer "does the subset *and its displayed shape* survive an input perturbation?" soundly but approximately.

A quantitative comparison, and what we did not do. Their supplementary implementation is not publicly downloadable, so we did *not* run their code. What we can do is compute the objective their algorithm computes — Def. 1 of Bassan et al. (2026): pin S at x, let the omitted features range over an ε_p-ball, require the decision to be preserved — exactly, on our own models. Cardinal minimality is a well-defined optimum, so the subset *sizes* below are the sizes their algorithm returns. At $d \le 13$ we obtain them by enumeration rather than by their sorted binary search; **their algorithmic contribution is reaching the same optimum in $O(\log d)$ verification queries, which our enumeration does not match and which is what matters at larger d.** Our timings are therefore not a speed claim. The comparison is on the clean protocol (canonical capacity, held-out split, five seeds, 25 instances per dataset, $G{=}10^6$) — the same *models* as Table 43, though on 25 of its 40 evaluation instances — and it is *minimal against minimal, signed against signed*: their cardinally-minimal subset against VIB-X's minimal explanation ($\varepsilon_{\text{sel}}{=}\varepsilon$), both certified with the signed, ball-localised bound of arm D. Their subsets are shown at the `full` level, so $e_i^{\mathcal{C}}{=}0$ and arms B and D coincide on their side.

Table 27: **The abductive objective computed on our models.** Clean protocol, five seeds, 25 instances per dataset, $G{=}10^6$. Both radius columns are arm D; on their side the subsets are shown at the `full` level, so arms B and D coincide. Empty subsets are scored as they certify, not dropped.

		minimal $\|S\|$		arm-D radius		their S	their S	our S
Dataset	d	theirs	ours	their S	our S	$r > 0$	empty	sufficient?
`abalone`	8	2.66	3.34	0.238	**0.388**	58%	0%	72%
`wine`	11	3.43	4.13	0.217	**0.301**	77%	0%	29%
`diabetes_reg`	10	1.37	3.11	0.196	**0.427**	64%	19%	90%
`heart`	13	2.12	7.95	0.062	**0.339**	22%	47%	88%

Reading. Their subsets are smaller, as they should be: the objective is exactly minimal, requires only that the *decision* survive rather than that a displayed surrogate stay within ε, and pins S rather than perturbing it. Two things the comparison is not: "minimal against minimal" pairs their *cardinally* minimal subset with our *greedy* minimal one, which is 1.2–3.8× larger and not claimed to be optimal; and both sides are scored at one absolute $\varepsilon{=}0.5$, which on `heart` is a tight tolerance against a logit spread of $\sigma_{\text{pred}}{=}2.24$ (App. A46 prices that effect for the backbone comparison). Neither favours one side over the other — both arms see the same ε on the same model — but both bound how far the numbers travel. The cross-evaluation is the informative column, and it has to be scored with the right bound: Bassan's Def. 1 constrains the *signed* reachable interval of the omitted set, which is nearly what arm B measures, so charging their subsets $\sum_{i \notin S} |f_i|$ (arm A) would score them against a condition their method does not ignore — under that bound they certify a positive radius on only 11% of instances, under the signed one on 55% pooled (58/77/64/22%). Like for like, our minimal explanation certifies the wider arm-D radius on all four datasets (4/4 on four seeds, 3/4 on one); the margin is 1.4–5.5× and is largest on `heart`. **The empty subsets are in that column**, scored as they certify rather than dropped: an empty set is a correct answer to their question — theirs is

empty on 47% of `heart` instances at ε_p=0.5 (the "empty" column) — and with nothing displayed the criterion reads $\sum_{\text{all } i} M_i(r) \le \varepsilon$, which arms B and D answer identically, so their side of this table is the same number under either. 13 of the 24 empty `diabetes_reg` subsets certify a *positive* radius, which is why the column is not simply zeroed. Dropping the empties instead would read 0.207 on `diabetes_reg` and 0.120 on `heart` — a margin of 1.4–2.8× — but that column would be a mean over 66 of `heart`'s 125 instances printed beside our mean over all 125, which is the two-population splice this paper spends its length on. Conversely our minimal subsets, 1.2–3.8× larger, are already sufficient in their sense on 29–90% of instances — the same question asked of the stability-selected explanation gives 42–97%. **We do not offer any of these numbers as evidence for our thesis**; their criterion never promised stability under perturbation of S, and the two objectives are incomparable in both directions. The stability-selected explanation (target radius 0.05) the runner also records is 3.6–8.8 features and is not shown in this table, because it is not minimal and a comparison against a cardinally-minimal subset must be. Cardinal minimality does not pin a *particular* subset either: where several are minimal, the enumeration returns the first in index order, so "their S" is one of possibly many optima and its certified radius is a property of that choice.

Where the objective binds. At small ε_p their problem is degenerate — if the omitted features cannot move far enough to change the decision, $S = \emptyset$ is already sufficient. Sweeping $\varepsilon_p \in \{0.05, 0.1, 0.25, 0.5\}$, mean $|S|$ goes $0 \to 0 \to 1.42 \to 2.66$ on `abalone` and $0.12 \to 0.27 \to 0.92 \to 2.12$ on `heart` — the terminal values are the table's *theirs* column, as they must be. We report at ε_p=0.5, matching the radius our own certificates use, so the two methods are compared on the same ball. Artifact: `results/abductive/abductive_comparison_seed42.json` and its seed 43–46 companions (`elasticnam/experiments/run_abductive_comparison.py`).

A39 Certification Controls: Verifier Tightness, Capacity, Depth

How far five seeds get us. Seed-to-seed dispersion is small (2.2% on the log scale) and seed 42 is the *least* favourable of the five, so a single-seed reading of it understates the effect. Two claims are stated per seed: $\ge 1.3\times$ holds on $8/9$ datasets in *all five* seeds — narrowly at the bottom: `concrete` clears it at 1.314 on one seed, a margin of 1% over a threshold the registration admits was written with a 2.06× result in hand — and $\ge 2\times$ on 3–$5/9$, seed 42 lowest. The one loss is also stable: `california` is below 1 in four of five seeds (0.966 geometric mean, range $[0.883, 1.031]$), so the reversal below is a property of that dataset, not an unlucky draw. Artifact: `results/robust_cert/seed_variance.json`, aggregated from `results/robust_cert/exact_cert_a_seed42.json` and `results/robust_cert/exact_cert_b_seed42.json` with seeds 43–46, which Fig. 11 also draws from. Those ten files carry §4's 1.87×; they are named by full path because the freshness gate finds cited artifacts by their `results/` path, and an abbreviation would leave them outside it.

Three ways the comparison could be unfair. Comparing a near-exact certificate against a CROWN *relaxation* invites three objections, and each has a control below, including the one that goes against us. **(i) Verifier looseness**: against sound α-CROWN the headline barely moves ($\mathbf{1.78}\times$ against 1.86× for vanilla CROWN on the same nine datasets, seed 42), and a PGD *upper* bound resolves it per dataset — $6/9$ real, $2/9$ unresolved, $1/9$ **reversed** (`california`, where the MLP provably certifies wider). **(ii) Capacity**: against a parameter-matched MLP the advantage *grows* to 2.35×. **(iii) Depth**: the fair parameter- *and* depth-matched cell certifies $\mathbf{1.86}\times$ $[1.77, 1.96]$ — a separate run, and a coarser one, at G=2×10^5 over $\beta \in \{0, 1\}$, which costs it 13% (1.65× against 1.87× on the same nine datasets). Between seeds both mechanism effects exclude 1; clustered by dataset only the depth effect survives, so we assert that alone and do not apportion the growth (one seed suggests 46%; five seeds span 13–62%).

The two predicates, reported apart. The headline pools two different certified quantities: on the six regression datasets the radius is the largest ε with worst-case $|\Delta f| \le \tau$=0.5 standardised target units, and on the three classification datasets it is the largest ε with no logit sign flip. Pooling them into one geometric mean is a convenience, not a claim that they are commensurable, so we give them separately: the five-seed geometric mean is 1.70× over the six regression datasets and 2.23× over

the three classification datasets. The advantage is therefore larger on the classification predicate, and a reader who cares about only one of the two should use that column rather than the pooled $1.87\times$.

The body summarises three controls in one paragraph; here is each in full.

(i) Verifier tightness. For each dataset we compute vanilla CROWN, α-CROWN (strictly tighter, still sound, same family (Xu et al., 2020; Wang et al., 2021)), and a multi-restart PGD *upper* bound on the MLP's *true* radius — an attack that succeeds proves the true radius is smaller. Together these bracket the MLP's true radius, so each dataset can be classified rather than aggregated. **Real** $(6/9)$: `concrete`, `bike`, `abalone`, `heart`, `breast_cancer`, `adult` — VIB-NAM's radius exceeds the attack bound, so no verifier improvement could close the gap. **Unresolved** $(2/9)$: `wine`, `diabetes_reg` — VIB-NAM lies between α-CROWN and the attack bound, so the gap *may* be looseness. **Reversed** $(1/9)$: `california`. Geometric mean against α-CROWN is $1.78\times$ versus $1.86\times$ against vanilla CROWN on the same nine datasets and the same seed, so the tighter verifier costs the headline 4% and does not explain it. Artifact: `results/robust_cert/verifier_tightness_seed42.json`.

(ii) Capacity. The protocol-matched baseline carries $\sim$30–$100\times$ fewer parameters than VIB-NAM, so we repeat against an MLP whose width is chosen to match the parameter count to $< 0.8\%$ (e.g. 1.088M vs. 1.092M on d=8), plus a quarter/double sweep. The advantage grows to geometric-mean $2.35\times$ over the five seeds that carry a parameter-matched arm; on seed 42's width sweep alone it is $2.23\times$ against $1.81\times$ for the small net *in that same run* (the pooled five-seed small-net figure is $1.87\times$), and `california` — the one loss — becomes a win at $1.21\times$. Artifacts: `results/robust_cert/exact_cert_a_seed42_b2capacity.json` and its `_b_` counterpart.

(iii) Depth, separated from capacity. The parameter-matched net above has three hidden layers against the protocol net's two, and CROWN's relaxation error compounds with depth, so parameters and depth are confounded in (ii). We added a depth-matched parameter-matched net (two hidden layers, same budget) and ran the full 2×2 cell on **five seeds** (42–46).

Table 28: The depth×capacity cell: five seeds (42–46), geometric mean and dataset-clustered 95% CI. This is the control that separates depth from parameters, and §5.1 quotes its $1.86\times$; it too shipped uncaptioned and unnumbered.

VIB-NAM's radius against	geo-mean (5 seeds)	95% CI
the small protocol-matched net	1.653	$[1.573, 1.736]$
parameter-matched, *depth-matched*	**1.864**	$[1.769, 1.964]$
parameter-matched, deeper (3 layers)	2.042	$[1.950, 2.139]$
effect of matching parameters (fixed depth)	0.887	$[0.854, 0.920]$
effect of depth (fixed parameters)	0.912	$[0.850, 0.979]$

Does the advantage survive conditioning on fit? It does, and the question had to be asked: a certified radius under $|\Delta f| \le \tau$ is monotonically *improved* by flattening the model — a constant predictor certifies $r = \infty$ — so a comparison between models of unknown relative fit measures nothing. Until this revision the `mlp_crown` blocks recorded radii, parameter counts and timings but no clean score, so the objection could not be checked even from the supplement. Both arms now record test R^2/AUROC in the same run, over all 45 (dataset, seed) cells.

The MLP fits better in $29/45$ cells, by $+0.039$ on average, so the confound is live rather than hypothetical. It is not what drives the result. Regressing $\log$(radius ratio) on the fit gap gives $\log \rho = 0.607 + 0.431\,\Delta_{\text{fit}}$, so the predicted ratio at *matched* fit is $\mathbf{1.83}\times$ against the unconditioned $1.87\times$, and leave-one-dataset-out puts the matched-fit estimate between $1.72\times$ and $2.01\times$.

That point estimate is less stable than a single fitted line suggests, and it must not be read as the confound being "under 2% of the effect". The first thing wrong with that phrasing is that there is no identified effect to take a percentage of: the pooled correlation is $r{=}+0.12$ over n=45, p=0.44. The second is that the fitted sign contradicts the split reported at the end of this paragraph — a positive slope on $\Delta_{\text{fit}} = (\text{MLP} - VIB - NAM)$ says our advantage *grows* where the MLP fits better, while the split says it shrinks ($2.04\times$ against $1.78\times$). Both cannot describe a real relationship, and the

resolution is that neither does. The pooled slope is an average over two predicates this appendix declines to pool a few paragraphs above. Refit within predicate it is $+1.20$ on the six regression datasets (matched fit $1.59\times$, n=30) and -8.43 on the three classification datasets (matched fit $2.16\times$, n=15); dropping `bike` alone flips the pooled slope to -2.46. `bike` carries the leverage — the additive model is 0.17 to 0.32 R^2 worse there and certifies 2.2 to $2.6\times$ wider — which is also why the unconditioned and conditioned estimates sit so close. Within the regression predicate the confound absorbs 13.5% of the log effect (on the linear scale the same ratio reads 7%, which is the category error behind "under 2%"). Treating the 45 cells as independent also contradicts the clustering this paper insists on elsewhere; a leave-one-dataset-out jackknife over the 9 clusters gives $\mathbf{1.82}\times$, 95% CI $[1.39, 2.38]$. That interval, not the regression's third digit, is what the control establishes: the estimate ranges over 1.59–$2.16\times$ across every subset we tried and excludes 1 in all of them, so the advantage survives conditioning on fit, but its magnitude is known to roughly a factor of 1.5, not to two decimals. One caveat the regression cannot address: ΔR^2 is a proxy for the wrong quantity. The mechanism at issue is smoothness, and a structurally additive model is *constrained* flatter than an MLP even at equal R^2, so matching fit does not fully match flatness. The correlation between the fit gap and the log ratio is $+0.12$, i.e. essentially none, and the advantage is if anything *larger* in the 16 cells where the MLP does not fit better ($2.04\times$ against $1.78\times$ where the MLP does). A flatter model certifying wider is a real mechanism; it is not this one.

Matched accuracy without a regression. The regressions above adjust for fit; they do not compare at equal fit. The β sweep records five VIB-NAM models per cell, so the dominance question can be asked directly: is there a model on the sweep that fits at least as well as the MLP (test score $\geq$ the MLP's) *and* certifies a larger mean radius? Such a fit-matched model exists in 19 of the 45 cells (7 of 9 datasets; none on `abalone` or `bike`, where every sweep model fits worse), and in all 19 it certifies more: geometric mean $2.05\times$, never below $1.03\times$. Choosing within the sweep by test score makes this a statement about what the sweep contains, not a deployable selection rule. In the other 26 cells no model of ours reaches the MLP's fit, and there the radius comparison is at unequal accuracy and says nothing about equal-accuracy robustness; the $1.87\times$ headline should be read as a comparison of the models as selected, which is how the body states it.

The deployable version. The $19/45$ above conditions on test scores and is a post-hoc existence result. A rule that looks at no test score — keep the sweep models whose *validation* score is at least the MLP's, take the one with the largest *validation* radius, abstain otherwise — selects a model in 15 of the 45 cells. On test data all 15 certify a wider radius than the MLP (geometric mean $2.24\times$), and 11 of the 15 are also at least as accurate on test; in the other 30 cells the rule abstains, because no model of ours matches the MLP's validation accuracy. Both analyses: `scripts/fit_pareto.py`, from the same artifacts.

Between seeds each interval excludes 1; clustered by dataset only the depth effect does (item (iii) above), so the capacity effect is suggested, not established, and the sentence that matching parameters "genuinely" shrinks what CROWN can prove is withdrawn. We report the fair cell's $\mathbf{1.86}\times$ as the honest number.

A claim we do not make. Apportioning the growth between the two mechanisms — "depth accounts for roughly 45% of the log-growth from $1.60\times$ to $1.95\times$" — is what a single seed suggests. Across five seeds the share is 46, 25, 62, 13 and 56% — mean 40%, range $[13, 62]$ — so the apportionment is not identified by this experiment. Between seeds both effects exclude 1; clustered by dataset only depth does. We make no claim about their relative size, nor that the parameter effect is non-zero. This is precisely the failure mode that motivated running seeds at all: the decomposition operated on differences of the same order as the between-seed variation, and only replication could reveal which of the two survived.

The cell uses $\beta \in \{0, 1\}$ and a 2×10^5-point grid rather than the headline's five-point sweep and 10^6 points, which removes the β our selection rule picks on 2–7 of the nine datasets depending on the seed ($6/4/2/7/4$ for seeds 42–46) and therefore handicaps VIB-NAM. Artifacts: `results/robust_cert/exact_cert_depthcell_seed42.json` (and `...seed43` through `...seed46`), `results/robust_cert/depthcell_variance.json` (`scripts/depth_cell_variance.py`).

A40 How the Frontier Scales with d

Bound: unsigned (arm A) throughout this appendix except where a column says arm D, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: the small-capacity $(128, 128)$ *encoder with in-sample validation (App. A48).*

The frontier results in the body are on $d \leq 13$ and the VIB-X demo of App. A30 tops out at d=30. Since the central claim is about explanation *size*, the honest question is whether the frontier survives at larger d or degrades to vacuity. It degrades, and we measure how — then App. A45 separates the two things that are degrading, because only one of them is the criterion's.

A single real dataset cannot separate "d is large" from "this model is dense", so we run both: `spambase` (d=57, the only dataset in our registry with $d \geq 50$) and a synthetic sweep that holds the number of genuinely informative features *fixed at six* while d grows, so any rise in $|S|/d$ is attributable to d alone.

Table 29: **The d-sweep.** Three seeds, small-capacity $(128, 128)$ encoder with in-sample validation, G=10^6; the synthetic cases hold the informative count at 6. Both arms at minimality; the build column is the wall-clock of the certifier plus error table in the sweep's own run. Memory is not in the table: the runner's `grid_megabytes` field is float32 arithmetic ($8d$ MB) the code does not use. The certifier stores float64 ($16d$ MB), and the measured peak with the error table is 4.0 GB at d=100 (App. A21), where the same construction took 299 s on a GPU shared with other jobs — wall-clock varies several-fold with contention.

		$\lvert S\rvert/d$ at ε_{sel}			arm A, minimal		arm D, minimal		build
Case	d	0.50	0.35	0.20	radius	cert.	radius	cert.	(s)
`breast_cancer`	30	66%	72%	79%	0.005	0.71	0.298	1.00	10
`spambase`	57	64%	69%	77%	0.000	0.20	0.092	1.00	20
synthetic (6 informative)	20	49%	58%	69%	0.014	0.21	0.466	1.00	6
synthetic (6 informative)	50	59%	65%	73%	0.000	0.00	0.495	1.00	17
synthetic (6 informative)	100	68%	73%	79%	0.000	0.00	0.498	1.00	37

Both arms are read at minimality (ε_{sel}=ε), the operating point the collapse is about; at ε_{sel}=0.35 *arm A certifies* every *instance in every row, with mean radius* $0.178/0.043/0.386/0.427/0.429$*, so the collapse is a statement about minimal explanations and not about the criterion at any budget. The arm-D radii are censored at* $r_{\max}$=0.5 *on* $21/0/73/95/97\%$ *of instances (same row order), so on the synthetic sweep they are lower bounds: what they establish is that arm D certifies every instance where arm A certifies none, not how far. Re-run at* $r_{\max}$=2*,* d=100 *gives a mean arm-D radius of* 1.51 *over the same three seeds, with* 59% *still at the ceiling.*

The explanation does not scale. With the informative count pinned at six, VIB-X selects 9.7 features at d=20 — already more than the truth — and 29.5 at d=50, 68.2 at d=100 (three seeds). It is not recovering the six; it is selecting a fraction of d that *increases* with d. The cause is structural rather than a defect of the greedy: acceptance requires $\sum_{i \notin S} |f_i(x_i)| + \sum_{i \in S} e_i^{\mathcal{C}_i} \leq \varepsilon$, and the first sum runs over the *excluded* features, so shape values that are individually negligible still accumulate past ε once there are enough of them. Any explainer using an additive residual budget inherits this; it is a property of the criterion, not of our search.

The certificate does scale. Everything else stays cheap and stays sound. Building the grid and error table costs 37 s and 4.0 GB peak resident memory at d=100 on the paper's G=10^6 grid (App. A21; both linear in $d \cdot G$); certifying one explanation costs 45 ms (seed 42); under arm D the *minimal* explanation certifies 0.498 on 100% of instances at d=100, where the unsigned bound certifies nothing at all — the "certifies nothing for $d \geq 50$" boundary is the unsigned bound's, not the certificate's. So the $O(d \cdot G)$ certificate is not the bottleneck, and a future explainer with a better size–d profile would inherit our machinery unchanged. A shared-basis head such as NBM (Radenovic et al., 2022), which preserves the per-feature one-dimensional structure both certificates need, is the natural place to look.

What we therefore claim. The stability–simplicity frontier is a tool for small-d or genuinely sparse models. At $d \gtrsim 50$ the stabilised explanation names most of the input and is not an explanation

in any useful sense, whatever radius it certifies. We state this as a scope boundary rather than as future work, because the sweep above shows it is a property of the objective. Artifact: `results/scale/scale_frontier_seed42.json` (`elasticnam/experiments/run_scale_frontier.py`).

A41 Is Fragility the Greedy's Fault? The MIP Optimum Says No

Bound: both, on the same selections of the same models — the explainer's unsigned acceptance test (arm A) and the signed, ball-localised residual it displays (arm D). Ceiling: $r_{\max}$=2.0*, where* 0–12% *of these radii are censored; the* $r_{\max}$=0.5 *replicate pins* 48–92% *of both* arm-D *columns at the ceiling (arm A there is at most* 4%*), which is enough to make those two agree, and it is reported as a separate measurement rather than as this one:* `--r_max` *widens the error table as well as the ceiling, so the two runs do not certify the same* $e_i^{\mathcal{C}}$*. Protocol: the small-capacity* $(128, 128)$ *encoder with in-sample validation (App. A48).*

The headline protocol first. The study below was run on the small in-sample protocol, one seed and 25 instances. Re-run under the protocol every headline number uses — canonical capacity, held-out validation, G=10^6, $r_{\max}$=0.5, 40 instances, five seeds — the greedy and the exact optimum select the same $|S|$ on every dataset, and the comparison reads:

Table 30: Greedy against the exact MIP optimum under the headline protocol (five-seed means; arm A / arm D on the same selections). The MIP solved every instance.

Dataset	$\lvert S\rvert$	greedy A / D	MIP A / D	MIP/greedy A / D
`abalone`	3.21	0.1536 / 0.3751	0.1448 / 0.3668	0.94 / 0.98
`wine`	4.15	0.0497 / 0.2929	0.0436 / 0.2878	0.88 / 0.98
`diabetes_reg`	3.10	0.0729 / 0.4335	0.0535 / 0.4215	0.73 / 0.97
`heart`	7.93	0.0090 / 0.3423	0.0141 / 0.3324	1.56 / 0.97

Under arm D the exact optimum certifies 0.97–0.98× the greedy in the five-seed means. Seed by seed that is not a stable ordering: the optimum's mean radius is at most the greedy's on 5/3/4/3 of five seeds, and the seed-paired difference is -0.008 $[-0.013, -0.004]$ on `abalone` and includes zero on the other three (widest $[-0.033, 0.009]$ on `diabetes_reg`). With no equivalence margin set in advance, what this supports is that the returned optimum does not certify more on average; it is not evidence that the two are equal, and — since the MIP minimises cost, not radius — it is not a test of the best radius at that cost, which is taken up below. Under arm A the small protocol's "the optimum is *more* fragile on all four datasets" does not replicate: on `heart` the optimum certifies 0.014 against the greedy's 0.009 on all five seeds. Both are near zero against arm D's 0.33, so the conclusion the body draws — no exact solver of the same cost fixes fragility, and the bound does — is unchanged; the stronger claim that exact optimisation makes it *worse* under both bounds is withdrawn for arm A. Artifacts: `results/mip_fragility/mip_fragility_clean_seed42.json` and seeds 43–46 (`elasticnam/experiments/run_mip_fragility_clean.py`, `scripts/mip_fragility_clean_summary.py`).

Every minimum-cost explanation, not one. A MIP returns one optimum, and optima can tie with very different radii: two features, one flat and one steep, each alone at the same cost, can differ a hundredfold (a reviewer's example). So we enumerated *every* minimum-cost explanation of each instance — solve for the optimal cost, then re-solve under that cost with a no-good cut on each assignment found, until infeasible — and certified each one (`elasticnam/experiments/run_mip_ties.py`; same models and protocol, five seeds; on all 800 instances CBC proved the program infeasible after the last cut, so every optimum was found — completion is read from the solver's proof, never from a time-out, and the enumerator is tested against exhaustive search; `results/mip_fragility/mip_ties_seed42.json` and seeds 43–46).

Table 31: All minimum-cost explanations, certified. "tied" is the share of instances with more than one optimum; "most" the largest number of optima on one instance. Radii: greedy / the best-radius optimum, five-seed means. "spread" is the largest ratio between the arm-D radii of two optima of

one instance. Last column: share certifying $r \geq 0.1$ for the best arm-A optimum / the greedy under arm D.

Dataset	tied	most	arm A	arm D	spread	$r \geq 0.1$
abalone	32%	7	0.1536 / 0.1500	0.3751 / 0.3820	46.3×	50% / 94%
wine	66%	59	0.0497 / 0.0506	0.2929 / 0.3358	327.7×	14% / 94%
diabetes_reg	42%	11	0.0729 / 0.0635	0.4335 / 0.4347	5.6×	21% / 99%
heart	72%	31	0.0090 / 0.0182	0.3423 / 0.3495	11.6×	2% / 92%

The reviewer's point holds instance by instance: optima tie on 32–72% of instances, and two optima of one instance can differ by up to 328× in arm-D radius. It does not change the average picture. Choosing the best-radius optimum lifts arm D by at most 15% (wine, $0.293 \to 0.336$), and under arm A it leaves the minimal explanation certifying $r \geq 0.1$ on 2–50% of instances, against 92–99% for the greedy's own explanation under arm D. What is established is therefore: on these four datasets, radius-aware tie-breaking among minimum-cost explanations recovers little, and far less than changing the bound. It is not a theorem about every program: a different cost, or selecting for radius directly, is a different explainer (`scripts/mip_ties_summary.py`).

The original study. The claim that minimality causes fragility is stated in terms of a greedy that stops the instant its budget is met, which invites the objection that a better search would find a small explanation that also leaves slack. If so, the right response would be a better solver rather than the frontier. We test it by certifying the *exact* optimum of the same program, using the MIP of §5.2, and a second variant whose cost weights make the feature count dominate so it minimises $|S|$ outright.

Table 32: **Greedy against the exact optimum, under both bounds.** One seed, 25 instances per dataset, small capacity, in-sample, $r_{\max}$=2.0 where 0–12% of these radii are censored. Each cell is arm A / arm D on the same selection.

	greedy (VIB-X)		MIP optimum		MIP, min $\lvert S\rvert$	
Dataset	$\lvert S\rvert$	A / D	$\lvert S\rvert$	A / D	$\lvert S\rvert$	A / D
abalone	2.76	0.129 / 0.726	2.76	0.065 / 0.658	2.76	0.059 / 0.671
wine	3.08	0.087 / 0.745	3.08	0.069 / 0.712	3.08	0.069 / 0.687
diabetes_reg	3.52	0.080 / 1.011	3.52	0.068 / 0.829	3.52	0.053 / 0.799
heart	7.92	0.031 / 0.530	7.92	0.026 / 0.520	7.92	0.020 / 0.501

Reading. The subset sizes are identical — on these instances the greedy already attains the optimal $|S|$, consistent with the ≤ 1.009 cost ratio of §5.2. What differs is the radius, and it goes the *wrong* way for the "better search" hypothesis: under the unsigned bound exact optimisation certifies a *smaller* radius than the greedy on **all four** datasets — by 1.99× on abalone, 1.27× on wine, 1.19× on heart and 1.17× on diabetes_reg. The mechanism is visible in the third pair of columns: at equal $|S|$ the optimiser is free to also pick cheaper functional forms, and a cheaper form carries a larger $e_i^{\mathcal{C}_i}$, which consumes budget that a perturbation would otherwise have had. Driving $|S|$ down outright (the last pair) is worse again, and on every dataset: 1.26–2.19× below the greedy under the same bound, heart included.

On this protocol the optimum is worse under *both* bounds (under the headline protocol above it is for arm D only). The same three explanations certify 0.50–1.01 under the signed bound against 0.020–0.129 under the unsigned one, and on every dataset under *either* bound the exact optimum certifies *less* than the greedy: 0.82–0.98× under arm D and 0.50–0.85× under arm A. So the search is not the cause under either bound — at equal $|S|$ optimising the same cost harder buys nothing and usually costs — and no solver substitutes for changing the budget, which is what the stability–simplicity frontier does. The direction is the same as the earlier run reported and the magnitude is not: there the arm-D columns agreed to 0.95–1.02× because 48–92% of both sat at $r_{\max}$=0.5, and a band between two columns pinned to the same ceiling is a property of the ceiling. That run is not a replicate of this one at a different ceiling. `--r_max` sets the bisection ceiling, the certifier's grid span (hence the knot spacing and the sound padding $L_i h$) *and* the margin the error table is computed over, so raising it inflates every $e_i^{\mathcal{C}_i}$ as App. A49 says it must. The evidence is that the arm-A radius moves on 99 of the 100 instances while only 1 ever sat at the 0.5 ceiling: what changed is the error table, not the ceiling. The selections are untouched — $|S|$ is identical on all 100. What does *not*

survive the wider ceiling is the arm-D claim we drew from the censored run, that there is little left for a solver to fix because the near-zero level is gone: at $r_{\max}=2$ the arm-D radii are 0.50–1.01 and still far from the ceiling, and the gap between the two arms (5.6–$17.2\times$ on the greedy column, 5.6–$25.4\times$ across all three) is the finding.

The mechanism. At equal $|S|$ the optimiser is free to pick cheaper functional forms, and a cheaper form carries a larger $e_i^{C_i}$, which consumes budget a perturbation would otherwise have had. One seed, 25 instances per dataset; a direction, not a measurement of its size. This is the evidence for that claim; the cross-evaluation against Bassan et al. (2026) in App. A38 is *not*, because their criterion is infeasible under ours at $r=0$ for reasons unrelated to minimality. Artifacts: `results/mip_fragility/mip_fragility_rmax2_seed42.json` (the table) and `results/mip_fragility/mip_fragility_seed42.json` (the censored replicate) (`elasticnam/experiments/run_mip_fragility.py`).

A42 Does the Frontier Transfer? EBM Says Partly

Bound: unsigned (arm A) throughout this appendix, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: the small-capacity $(128, 128)$ *encoder with in-sample validation (App. A48).*

We assert twice in the body that both certificates apply to any additive model. That was a claim about a construction, not a demonstration, so here is the demonstration — on the model a deployer is most likely to already have.

Why EBM, and why the certificate is exact there. EBM is the strongest accuracy baseline in Table 24 and the model we report as far more robust empirically (App. A57). Its main effects are *piecewise constant*, so the interval extremum is not approximated at all: build the grid with two knots just inside each bin and one on each cut, read values by exact bin lookup rather than by interpolating the grid, and the windowed extremum *contains* the true extremum with padding exactly 0. It is tight, not equal: a window whose end lands on a cut also admits the neighbouring bin's knot, which widens the interval in the conservative direction. This is not the $L_i h$ argument of Thm 3 with a small constant — a step function has no finite Lipschitz constant, and using `analytic_shape_lipschitz` here would be wrong rather than loose. `tests/test_ebm_cert.py` checks the intervals against brute force on every (instance, feature, radius) triple in all four datasets, including probes placed a nanometre inside each bin cut – the case an interval certifier is most likely to get wrong.

Table 33: **EBM under both certificates.** Five seeds, 40 instances per dataset, $\varepsilon=0.5$, EBM with `interactions=0`; the interval is exact for piecewise-constant effects, so no Lipschitz padding enters. Arm A throughout, as this appendix's banner says.

	$\varepsilon_{\text{sel}}=\varepsilon$ (minimal)				$\varepsilon_{\text{sel}}=0.35$	
Dataset	$\lvert S\rvert$	budget used	radius	certif.	radius	certif.
abalone	3.18	79%	0.181	1.00	0.352	1.00
wine	3.45	85%	0.060	0.99	0.207	1.00
diabetes_reg	2.77	91%	0.072	0.98	0.339	1.00
heart	8.42	85%	0.159	0.99	0.323	1.00

Five seeds, 40 instances per dataset, $\varepsilon=0.5$, EBM with `interactions=0`. Artifact: `results/explanation_stability/ebm_frontier_seed42.json` and its four sibling seeds (`elasticnam/experiments/run_ebm_frontier.py`).

Both transfer, once a unit bug is out of the way. Tightening ε_{sel} buys 1.9–$4.7\times$ radius for 0.7–1.1 extra features, and EBM's minimal explanations spend 79–91% of the budget where ours spend 87–98%. So the budget-exhaustion mechanism is a property of the acceptance criterion, not of our architecture, which is what we claimed and had not shown. Units matter here: our trainer standardises the regression target, and an EBM arm that does not sees the same ε as $0.5\sigma_y$ on one side and $0.006\sigma_y$ on `diabetes_reg` on the other, which reads as a spend of 16–91% and a fragility that does not transfer. With matched units the outlier disappears.

The difference is what the spent budget leaves. A histogram's worst-case surrogate error is smaller than a 10^6-knot neural shape function's, so the same 85% consumption leaves EBM a usable radius and leaves us almost none: on `heart` its *minimal* explanation certifies 0.159 against our 0.014, and its stabilised one matches ours, 0.323 against 0.324.

What this costs the paper, stated plainly. Our title sentence is narrower than we wrote it. "Minimal explanations are fragile" holds in proportion to how much of the error budget a model class spends reaching feasibility, and a 10^6-knot neural shape function spends much more of it than a histogram does. That does *not* mean a practitioner wanting a certified-stable explanation today should fit an EBM: the comparison rests on the unsigned criterion, which this appendix is measured under throughout, and under the signed, ball-localised bound the gap it rests on collapses to 1.27–$1.41\times$ on the clean protocol (App. A46). We do not read the *ordering* there — EBM keeps `heart` — so what the collapse removes is the reason to switch, not the ranking. What survives is the narrower statement that an EBM's certificate needs no padding at all, which is a property of the shape family and not of the radius it leaves. What survives for VIB-NAM is that the frontier is visible on it, that the certificate extends to continuous shape functions at all, and that the diagnosis of *why* minimality is fragile — the budget, not the search — is architecture-independent.

A43 Form Fidelity: The Budget Term We Did Not Price

The acceptance condition $\sum_{i\notin S} M_i(r) + \sum_{i\in S} e_i^{\mathcal{C}_i} \le \varepsilon$ has two terms that consume budget, and VIB-X spends both. After choosing S it walks each selected feature down to the cheapest functional form that still fits, which raises $e_i^{\mathcal{C}_i}$ and eats slack a perturbation could otherwise have occupied. Our headline priced the frontier in extra *features* only, so this control asks what the other term is worth.

The protocol is the frontier runner's, with one line changed: after VIB-X returns its selection we re-certify the *same* subset with every form left at `full` ($e_i^{\mathcal{C}_i}{=}0$). $|S|$ is identical by construction; only what is displayed for each selected feature changes, from a fitted constant/linear/quadratic to the shape function itself. Both bounds are certified on the same selections of the same models in one run, seeded per (dataset, seed), on the clean protocol. **The ceiling matters here**: at $r_{\max}{=}0.5$ the arm-D greedy radius is already censored on up to 62% of instances, so a `full`-form gain cannot register; the table is therefore $r_{\max}{=}2.0$, where the arm-D greedy radius is censored on 0–4.5% and the `full`-form radius — the numerator — on 0–6.0%, so these gains are ratios of capped means on two datasets, not bounds (App. A49) (censored at $r_{\max}{=}0.5$ the arm-D gain reads 1.01–$1.05\times$).

Table 34: **Form fidelity: the same subsets shown faithfully.** Clean protocol, five seeds, $r_{\max}{=}2.0$; S is held fixed and every feature is restored from its fitted form to the shape function itself, so the only thing that changes is $e_i^{\mathcal{C}}$.

		unsigned (arm A)			signed (arm D)		
Dataset	$\lvert S\rvert$	greedy	`full`	gain	greedy	`full`	gain
abalone	3.21	0.188	0.352	$1.88\times$	0.600	0.672	$\mathbf{1.12}\times$
wine	4.15	0.054	0.067	$1.24\times$	0.323	0.329	$\mathbf{1.02}\times$
diabetes_reg	3.10	0.070	0.203	$2.89\times$	0.740	0.782	$\mathbf{1.06}\times$
heart	7.93	0.024	0.160	$\mathbf{6.75}\times$	0.454	0.479	$\mathbf{1.06}\times$

The form axis is the same artifact as the other two. The price of a simplified form is $e_i^{\mathcal{C}_i}$ — an unsigned supremum over the error table's whole tabulated range — and that is precisely the term arm D signs and localises to the ball (App. A49). Setting every form to `full` drives $e_i^{\mathcal{C}_i}$ to zero, so under the unsigned bound it removes a large, badly-priced quantity and the radius jumps 1.2–$6.7\times$. Under arm D that quantity was already priced correctly, and removing it buys 1.02–$1.12\times$. Read from the unsigned column alone, form fidelity looks like the cheaper axis and the frontier's "0.8 extra features" looks like an overstatement of what the subset buys; **neither survives arm D.**

Taken with App. A46, every structural quantity this paper priced behaves the same way once the bound is tightened: the backbone spread 1.1–$11.1\times \to 1.27$–$1.41\times$, the form axis 1.2–$6.7\times \to 1.02$–$1.12\times$, the feature knob 3.5–$14.9\times \to 1.4$–$1.6\times$ — each pair measured on one population at one ceiling, the knob included: both of its ends are the $r_{\max}{=}2.0$ run (at $r_{\max}{=}0.5$ its arm-A end reads

2.5–31.7× off a censored population). That is not three coincidences. It is one fact about $e_i^{\mathcal{C}}$ and the excluded sum, observed three times, and it is the paper's title.

Five seeds (42–46), up to 40 instances per dataset (instances where VIB-X returns an empty selection have nothing to certify and are dropped; each row records its own count), ε=0.5, G=10^6 knots — the same density §4 certifies at. Artifact: `results/explanation_stability/form_fidelity_clean_rmax2_seeds42-46.json` and its $r_{\max}$=0.5 companion `results/explanation_stability/form_fidelity_clean_seeds42-46.json`.

Consequences for the frontier. Under the unsigned bound the form axis looks like the cheaper one — 1.2–6.7× of radius for zero extra features — and read that way the "0.8 extra features" price would attribute to the subset a gain that belongs to the forms. Under arm D the form axis is worth 1.02–1.12× and the feature knob 1.4–1.6× uncensored, so the frontier's price is stated in the right currency after all, and the two-dimensional frontier (subset size against form fidelity) has one axis that matters. One consequence does survive: the claim in App. A41 that "no improvement in search can substitute for changing the budget" is about a search that minimises a cost with a positive complexity weight and therefore prefers cheaper forms; it does not explore the form direction, and under the unsigned bound that direction is where the radius is.

Form fidelity is not free. It is priced in a currency $|S|$ does not measure: a `full` form shows a curve where a `linear` form showed a slope. Whether a reader prefers nine linear summaries or eight curves is not something $|S|$ answers, and we do not claim the `full`-form explanation is simpler. On `heart` the minimal explanation certifies 0.024 under the unsigned bound and 0.454 under arm D before either axis is spent; the fragility result stands for the former, which is the bound the paper argues against.

A44 Does the Frontier Survive a Clean Validation Split?

Bound: unsigned (arm A) throughout this appendix, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: the small-capacity $(128, 128)$ *encoder with a* held-out *validation split — one of the two repairs of App. A48, applied alone.*

Eight explanation runners pass `Xtr[:200]` as the validation set (App. A4), so early stopping in every frontier number is in-sample. The question a disclosure cannot answer is whether that changed the result. `run_stability_frontier.py --holdout_val 0.2` answers it: the training rows are partitioned 80/20 before the model is built, the scaler is refitted on the reduced training rows (fitting it on the union would trade one leak for another), and everything downstream — grid, evaluation points, ε, seeds — is held fixed. Table 35 is the result over five seeds.

Table 35: Canonical (in-sample validation) versus a held-out 80/20 validation split, five seeds, G=10^6. Budget and certifiable fraction are at ε_{sel}=ε=0.5; $|S|$ and radius are at ε_{sel}=0.35.

	rows		budget used (%)		certifiable		$\lvert S\rvert$		radius	
Dataset	train	val	canon.	held-out	canon.	held-out	canon.	held-out	canon.	held-out
`abalone`	2338	585	87.4	87.2	0.995	0.990	3.31	3.21	0.383	0.373
`wine`	3637	910	93.2	94.3	0.945	0.950	4.74	4.86	0.326	0.328
`diabetes_reg`	247	62	94.5	95.3	0.935	0.940	3.82	3.85	0.409	0.395
`heart`	151	38	98.3	98.8	0.925	**0.880**	9.10	8.74	0.324	**0.237**

What survives. Every claim the frontier carries. The minimal explanation still consumes almost the whole budget (87.2–98.8% against the 87.4–98.3% of the arm this one differs from by a single change), the certified radius still rises monotonically as ε_{sel} tightens on all four datasets, and the price in features is the same to within 0.43 features anywhere on the grid. The largest radius change anywhere in the 4×5 grid is 0.087, and it is the one cell discussed next; every other cell moves by at most 0.024.

One dataset moves, and in the leak's direction. `heart` is the dataset where the defect was worst — its training split is 189 rows, so the "validation" set was the whole of it and early stopping was

entirely in-sample — and it is the only dataset whose certifiable fraction at $\varepsilon_{\text{sel}}{=}\varepsilon$ falls materially, from 0.925 to 0.880 (nine instances in two hundred; `abalone` loses one), with the mean radius there dropping from 0.014 to 0.008. Its headline frontier point also softens, $0.32 \to 0.24$ at $\varepsilon_{\text{sel}}{=}0.35$. The correct reading is that `heart`'s minimal explanations are *more* fragile than we reported, not less, so the leak was flattering the arm we argue against; the frontier's own claim — that spending selection budget buys radius — is unaffected and on `heart` becomes slightly stronger ($0.008 \to 0.237$, a $30\times$ gain, against $0.014 \to 0.324$, $23\times$).

Which protocol each appendix uses. On the clean protocol (canonical capacity, held-out split): the body's frontier (App. A48), the four certification arms (App. A49), the identity certificate (App. A53), the across-backbone headroom (Table 38), the form-fidelity control (App. A43) and the formal-XAI comparison (App. A38). Still on the small-capacity $(128, 128)$ encoder with in-sample validation, and saying so in their first line: the budget-consumption and frontier tables of App. A54, the set-stability comparison, the MIP-fragility and certificate-1 controls, the EBM frontier (App. A42), the d-sweep and its L_2 companion (Apps. A40, A45) and the calibration control of Table 37. The two protocols agree wherever both exist except on `heart`, whose minimal radius is the most protocol-sensitive quantity in the paper (0.014 in-sample small, 0.070 in-sample canonical, 0.009 clean) — which is why the body quotes the clean one everywhere. Artifacts: `results/explanation_stability/valholdout_frontier_G1e6_seed42.json` and its seed 43–46 companions.

A45 Which Half of the d-Collapse Is the Criterion, and Which Is the Ball?

Bound: unsigned (arm A) throughout this appendix, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: the small-capacity $(128, 128)$ *encoder with in-sample validation (App. A48).*

Summary (moved from the body). As d grows the explanation names more features *and* certifies less radius. The selection condition sums over $d - |S|$ excluded features, so $|S|$ must grow whatever ball is certified later — any framework on a summed residual budget inherits that half. The threat model is the other: under $\|\delta\|_\infty \le r$ all $d - |S|$ excluded coordinates may sit at their worst case at once, while under $\|\delta\|_2 \le r$ they compete for one budget, which we certify exactly. Where the L_∞ certificate has collapsed the same explanations certify $\mathbf{1.7}$–$\mathbf{2.5}\times$ more radius off the ceiling (`breast_cancer`, `spambase`) and little extra where it was not binding, so the ratio tracks collapse more than d and is under a third of it on a log scale. Both sides are arm A, so this separates the threat model from the criterion, not from the bound (App. A45, the $r_{\max}{=}2$ run it calls the measurement).

App. A40 reports two things together and they are not the same thing. As d grows, (a) the explanation names a larger fraction of the features, and (b) the certified radius falls. We can separate them, and only one is fundamental.

The criterion decides (a), and this is an identity, not a measurement. The selection condition $\sum_{i \notin S} |f_i(x_i)| + \sum_{i \in S} e_i^{\mathcal{C}_i} \le \varepsilon_{\text{sel}}$ is evaluated *at* x, with no perturbation in it, so no ball appears anywhere in it: the runner selects S once and hands the same list to both certificates. $|S|/d$ agreeing under the two threat models below is therefore guaranteed by construction and we should not have offered it as evidence. The substantive point stands on its own: the excluded-feature sum has $d - |S|$ terms, so once d is large enough that the tail of small $|f_i|$ accumulates past the budget, $|S|$ must grow whatever ball we later certify. **This half does not go away.**

The threat model decides (b). Condition (2) adds $\sum_{i \notin S} M_i(r)$, and under $\|\delta\|_\infty \le r$ every one of those $d - |S|$ coordinates may sit at its own worst case at once — a perturbation of total energy $r\sqrt{d}$. Under $\|\delta\|_2 \le r$ the same coordinates have to be bought out of one budget, so the sum is carried by a few terms rather than by d of them. We certify that exactly: maximising $\sum_{i \notin S} M_i(r_i)$ subject to $\sum_i r_i^2 \le r^2$ is a multiple-choice knapsack, and its LP relaxation — the same Dyer–Zemel structure as the VIB-X optimality bound — is a solver-free upper bound. Soundness needs one more step: M_i is tabulated on a finite ladder of radii, so a rung is priced at the *lower* neighbour's cost and credited with the *upper* neighbour's value, which dominates any continuous allocation because M_i is

non-decreasing. `tests/test_l2_stability.py` checks the result against $\sim$900 perturbation directions per instance.

Table 36: The same explanations, certified over an L_∞ and an L_2 ball. Three seeds; ε=0.5, $r_{\max}$=0.5. $|S|/d$ is at ε_{sel}=0.35 and is the same under both balls. Synthetic rows hold the informative count fixed at 6.

			ε_{sel}=0.35			ε_{sel}=0.20		
Case	d	$\lvert S\rvert/d$	L_∞	L_2	ratio	L_∞	L_2	ratio
synthetic d=20	20	58%	0.386	0.399	1.03×	0.476	0.485	1.02×
synthetic d=50	50	65%	0.427	0.440	1.03×	0.491	0.496	1.01×
synthetic d=100	100	73%	0.429	0.444	1.04×	0.496	0.498	1.01×
`breast_cancer`	30	72%	0.178	0.280	**1.57**×	0.375	0.419	1.12×
`spambase`	57	69%	0.043	0.109	**2.52**×	0.119	0.253	2.14×

Read the ratios, not the radii — and they track collapse, not d. The L_2 ball of radius r is a *subset* of the L_∞ ball of the same radius, so the two radii are not comparable as numbers and L_2 being larger is not a finding. What is a finding is *where* the gap opens, and it is not where the $r\sqrt{d}$ argument alone would predict. The table above runs at $r_{\max}$=0.5, which censors the synthetic rows; re-run uncensored at $r_{\max}$=2.0 (three seeds, G=10^6) the ratios at ε_{sel}=0.35 are 1.02 (d=20), 1.09 (d=50) and 1.17 (d=100) on the synthetic rows, whose L_∞ radii are all above 1 (1.05/1.24/1.15), against 1.74 (`breast_cancer`, d=30) and 2.52 (`spambase`, d=57), whose L_∞ radii are 0.18 and 0.04. The gap opens where the L_∞ certificate has collapsed; where it has not, dimension alone moves the ratio only mildly, far below what "L_∞ lets energy grow as $r\sqrt{d}$" would predict. The honest mechanism is therefore narrower: the L_∞ sum is dominated by many small M_i terms only once the budget is tight relative to them, and dimension is one of several ways to get there. Artifact: `results/scale/scale_frontier_l2_rmax2_seed42.json` and its seed 43–44 companions.

And it removes under a third of the collapse, not most of it. At ε_{sel}=0.35 and off the ceiling ($r_{\max}$=2, three seeds), going from synthetic d=20 to `spambase` costs 24.3× of radius under L_∞ and 9.8× under L_2 — so on a log scale the choice of ball accounts for **29**% of the collapse and seven tenths of it survives. (At $r_{\max}$=0.5, where the synthetic rows are censored, the same arithmetic reads 8.9×, 3.7× and 41%; the uncensored run is the measurement.) "Partly", not "largely", is what the numbers support.

We do not claim L_2 is the right ball; which ball is right depends on whether measurement errors are coordinated. We claim that the choice has been treated as a formality in this literature and is instead a first-order determinant of whether a minimal explanation is certifiable at scale. Artifact: `results/scale/scale_frontier_l2_seed42.json` and its seed 43–44 companions (`elasticnam/experiments/run_scale_frontier.py --with_l2`).

A46 Explanation Headroom Looked Like a Property of the Model Class; It Is the Bound

Protocol: both. Table 38 and every number the body quotes are on the clean protocol (canonical capacity, held-out split); Table 37 and its calibration control are the small-capacity in-sample arm the study began on. Each statement below says which.

If the budget-exhaustion mechanism is universal but the radius it leaves is not, then the useful question for a practitioner is not which backbone is most accurate but which leaves the most room once a minimal explanation has been paid for. The certificate answers that directly, in milliseconds per instance. This appendix measures the answer under both bounds, on the protocol the body quotes (canonical capacity, held-out validation) and on the small-capacity in-sample arm the study began on. The finding is the same on both: an 11× spread under the unsigned bound collapses to 1.3–1.4× under arm D on the clean protocol (1.3–2.3× in sample), so what looked like a property of the model class was the bound. The *ordering* is not the same on both: in sample our backbone leads 4/4 under arm D, on the clean protocol EBM keeps `heart`, and we withdraw the reversal accordingly.

Three additive backbones, certified by the *same* criterion at the same ε=0.5: our neural shape functions, an Agarwal-style plain NAM (certified by the same analytic operator-norm bound, extended to that class here — ReLU is 1-Lipschitz and eval-mode dropout is the identity, so the product of the linears' operator norms is certified), and EBM, whose piecewise-constant effects need no padding at all. Five seeds, 40 instances per dataset.

Table 37: Budget consumed by the minimal explanation, and the certified radius left afterwards. Five seeds. **This table is the *uncalibrated* arm** — one shared ε for all three backbones; the *calibrated* arm, where each backbone's ε is priced as the same fraction of its own output spread, is in the same artifact under `calibrated` and disagrees on who leads `heart` at 0.35 ($4/4$ for us uncalibrated, $3/4$ calibrated). **Protocol: the small-capacity** $(128, 128)$ **encoder with in-sample validation**, and the unsigned bound (arm A) throughout — both the capacity/validation defects of App. A48 are present here. Table 38 repeats the comparison on the clean protocol under both bounds, and the body quotes that one.

	budget used at ε_{sel}=ε		certified radius	
Dataset (d)	VIB-NAM / plain NAM / EBM	range	minimal	at 0.35
`abalone` (8)	87.1 / 88.3 / 79.1%	9 pts	0.133 / 0.095 / **0.181**	0.383 / 0.331 / 0.352
`wine` (11)	92.7 / 90.3 / 85.1%	8 pts	0.058 / 0.049 / **0.060**	0.326 / 0.259 / 0.207
`diabetes_reg` (10)	94.2 / 95.0 / 91.5%	4 pts	0.067 / 0.035 / **0.072**	0.409 / 0.301 / 0.339
`heart` (13)	98.1 / 95.3 / 84.9%	13 pts	0.014 / 0.041 / **0.159**	0.324 / 0.239 / 0.323

Under the unsigned bound, at minimality the model class appears to decide. Every backbone spends 79–98% of ε on its minimal explanation, so no architecture escapes the stopping rule. What that leaves differs sharply: on `heart` the unsigned criterion certifies 0.159 for EBM, 0.041 for a plain NAM and 0.014 for ours — an 11× spread at identical ε, larger than any accuracy difference among the three (Table 24); EBM wins $4/4$ here, and $3/4$ on the clean protocol below. The ordering tracks how large a worst-case surrogate error each shape family admits, which is why a histogram wins and a 10^6-knot neural path loses — and that term is exactly what arm D signs and localises, which is why most of the spread does not survive it (below).

First, is the spread an artifact of an uncalibrated ε? It would be easy for it to be. ε is an absolute constant in output units and the budget spent is literally $\sum |f_i| + \sum e_i$, which scales with shape magnitude. The regression arms all standardise y, and their output spreads agree to within 11%; the classification *logit* scale is not standardised and does not agree — on `heart` we measure 2.92 (ours) / 3.48 (plain NAM) / 2.19 (EBM), so at a shared ε EBM is certified against a budget 1.59× looser than the plain NAM and 1.33× looser than ours. On this paper's own `heart` frontier a 30% budget change moves the radius nearly 20×, so a miscalibration of that size could manufacture the whole effect. We therefore re-ran every cell with ε priced as a fixed fraction of each backbone's own output spread (`--calibrate_eps`), making the effective tolerance $\varepsilon/\sigma_{\text{pred}}$ equal to 0.5 for all three.

It survives. At matched tolerance the `heart` spread is **11.3**× against 11.4× uncalibrated, and EBM is ahead on $4/4$ either way. The headroom difference is a property of the shape families, not of the units.

Second, is it just that the wider-radius model fits worse? A certified radius under $|\Delta f| \le \tau$ is monotonically *improved* by flattening the model — a constant predictor certifies $r = \infty$ — so the comparison means nothing between models of unknown relative fit, so the artifact records accuracy in the same runs: test R^2/AUROC for VIB-NAM / plain NAM / EBM is 0.590/0.597/0.573 (`abalone`), 0.265/0.298/0.323 (`wine`), 0.506/0.501/0.536 (`diabetes_reg`) and 0.9268/0.9177/0.9218 (`heart`). The largest spread is 0.058 and on `heart` — the row carrying the 11× — the gap is **0.009** and *our* model fits best while certifying narrowest, which is the opposite of the confound. Across the four datasets the widest-radius backbone (EBM every time) ranks $3/1/1/2$ on accuracy, so the radius ordering is not the fit ordering.

Then spend the selection budget, and it closes. At ε_{sel}=0.35, the operating point this paper argues for, the across-backbone spread collapses from 11× to **1.2–1.7**× and no backbone dominates. Both arms say so and they disagree on who leads `heart`: on the *calibrated* arm ours leads $3/4$ and EBM

takes `heart` (0.374 against our 0.358); on the uncalibrated arm of Table 37 above, ours leads $4/4$, `heart` included (0.324 against 0.323), and the spread is 1.2–1.6×.

And the spread does not survive the bound. It was the bound. This appendix explains the spread by saying the ordering "tracks how large a worst-case surrogate error each shape family admits, which is why a histogram wins and a 10^6-knot neural path loses". That term — $e_i^{\mathcal{C}}$, an unsigned supremum over the error table's whole tabulated range — is exactly what arm D signs and localises to the ball (App. A49), so the backbone with the largest e_i has the most to gain and the histogram the least. Re-run on the *clean* protocol (benchmark capacity, held-out validation — the protocol the body quotes; the table above is the small-capacity in-sample arm), recording both bounds on the same selections of the same models, five seeds:

Table 38: Mean certified radius of the *minimal* explanation, by backbone, under both bounds on the same models. Clean protocol (canonical capacity, held-out validation), five seeds, ε=0.5, $r_{\max}$=2.0, where at most 4.5% of instances are censored (the $r_{\max}$=0.5 replicate is in the text). Test R^2/AUROC per backbone: 0.593/0.579/0.583 (`abalone`), 0.308/0.304/0.302 (`wine`), 0.502/0.509/0.538 (`diabetes_reg`), 0.921/0.933/0.925 (`heart`), so no arm is certifying a flatter model. The bracketed interval resamples the five seeds and recomputes the printed ratio (4000 draws, percentile), so it is that number's own uncertainty rather than a different estimator's. The arms separate on `diabetes_reg` and `heart` — the two datasets with a large unsigned spread — and on `abalone` and `wine` the signed spread is the *larger* of the two, which is what a claim about the bound rather than the class predicts: where there is no order-of-magnitude gap to explain, there is nothing for the bound to close.

	unsigned (arm A)				signed (arm D)			
Dataset	ours	NAM	EBM	spread [95%]	ours	NAM	EBM	spread [95%]
`abalone`	0.188	0.172	**0.199**	1.2× [1.1, 1.3]	**0.600**	0.473	0.443	1.35× [1.22, 1.48]
`wine`	0.054	**0.061**	0.057	1.1× [1.0, 1.4]	**0.323**	0.317	0.254	1.27× [1.23, 1.39]
`diabetes_reg`	0.070	0.041	**0.120**	2.9× [2.3, 3.7]	**0.740**	0.544	0.531	1.39× [1.05, 2.02]
`heart`	0.024	0.059	**0.262**	11.1× [7.9, 16.7]	0.454	0.366	**0.514**	1.41× [1.18, 1.80]

The collapse survives; the reversal does not. On the clean protocol the unsigned spread at minimality is 1.1–11.1× uncensored (1.1–21.6× at $r_{\max}$=0.5, where `heart`'s minimal radius is 0.009 against EBM's 0.195), and under arm D it is **1.27–1.41**× (1.06–1.18× censored). Those are ratios of seed means; *within* a seed the arm-D spread runs 1.13–2.56× against the unsigned 1.03–22.4×, and the paired intervals in the table separate the arms on the two datasets that have a gap to explain. So the order-of-magnitude headroom gap *is* the bound, which is this appendix's finding and the paper's title. But the ordering does *not* reverse: EBM keeps `heart` under arm D (0.514 against our 0.454), and our backbone leads $3/4$ datasets uncensored ($2/4$ censored, $4/4$ at ε_{sel}=0.35), per seed 2/5–5/5. **We withdraw the claim that the ordering reverses**, which the small-capacity in-sample arm supported and this one does not. To be concrete about what we are declining to claim: on the clean protocol our backbone leads $0/4$ under arm A and $3/4$ under arm D, which is a change of leader — but the count moves to $2/4$ censored, $4/4$ at ε_{sel}=0.35 and 2/5–5/5 per seed, so it is not a finding, and we do not present it as one anywhere. What survives is the negative statement: at minimality the backbone is worth under 1.5× between seed means on this protocol — 2.56× within a seed, 2.3× in the small-capacity arm — against up to 11× under the bound that does not, so "fit a histogram" is not advice the spread supports (§5.3 says the same) and no backbone choice is worth making on these grounds. Artifacts: `results/backbone_headroom/backbone_headroom_clean_rmax2_seeds42-46.json` and `results/backbone_headroom/backbone_headroom_clean_seeds42-46.json` (`--signed --canonical --holdout_val 0.2`); the small-capacity in-sample replicates are `results/backbone_headroom/backbone_headroom_signed_seeds42-46.json` and `results/backbone_headroom/backbone_headroom_signed_rmax2_seeds42-46.json`. Both are named in full because the manifest, the reproduction index and the freshness gate find cited artifacts by their `results/` path: a basename alone is invisible to all three.

So "which additive model leaves the most room for a perturbation" is a question whose answer depended on how we asked it.

Reading the $11\times$ off the minimal column and telling a deployer to fit an EBM would take a recommendation from the one operating point the rest of the paper argues nobody should deploy, under the bound the title rejects. The honest statement is narrower, and it is the same on both protocols: at minimality the backbone is worth 1.27–$1.41\times$ on the clean protocol and 1.3–$2.3\times$ in the small-capacity in-sample arm, against up to $11\times$ unsigned; spend selection budget instead and the backbone stops mattering at all, so you may keep the model you wanted for other reasons. *Which* backbone leads is not the same on both — EBM keeps `heart` on the clean protocol while our backbone leads $4/4$ in sample — so we withdraw the ordering and read only the collapse. Four datasets, margins under $2.3\times$, no preregistration: what the appendix establishes is that the $11\times$ is gone, and nothing replaces it. Artifacts: the clean numbers from `results/backbone_headroom/backbone_headroom_clean_rmax2_seeds42-46.json`, the in-sample ones from `results/backbone_headroom/backbone_headroom_signed_rmax2_seeds42-46.json`, and the uncalibrated/calibrated pair of Table 37 from `results/backbone_headroom/backbone_headroom_seeds42-46.json` (`elasticnam/experiments/run_backbone_headroom.py`).

A47 Does the Frontier Survive at the Benchmark's Own Capacity?

Bound: unsigned (arm A) throughout this appendix, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: the benchmark's canonical $(256, 256, 128)$ *capacity with* in-sample *validation — the other repair of App. A48, applied alone.*

The explanation runners train a $(128, 128)$ encoder with $d_z{=}16$ for 60 epochs, while Table 24 reports a $(256, 256, 128)$ encoder with $d_z{=}32$ trained for 300 (App. A4). That was not a configuration choice: `AdditiveExactCertifier` and `ErrorTable` each evaluated the shape function on all 10^6 grid knots in a single forward pass, which allocates past a 24 GB card at the canonical width, so the certificate could not be built for the model the benchmark reports. Blocking that evaluation (`_eval_shape_on_grid`) removes the obstacle — a shape function is pointwise, and the residual float32 kernel-choice disagreement is 3.6×10^{-7} against a certified padding of $\sim 3 \times 10^{-4}$ — and `run_stability_frontier.py --canonical` then rebuilds the whole frontier there. Five seeds, same grid, same evaluation points.

Table 39: The frontier at the explanation runners' $(128, 128)$ capacity (sm.) and at the benchmark's canonical $(256, 256, 128)$ (can.). Five seeds, $\varepsilon{=}0.5$. $|S|$, budget and the first radius column are at $\varepsilon_{\text{sel}}{=}\varepsilon$; the second radius column is at 0.35; "gain" is their ratio.

		$\lvert S\rvert$		budget (%)		r minimal		r at 0.35		gain	
Dataset	d	sm.	can.	sm.	can.	sm.	can.	sm.	can.	sm.	can.
`abalone`	8	2.50	3.50	87.4	85.9	0.133	0.148	0.383	0.359	$2.9\times$	$2.4\times$
`wine`	11	3.52	4.12	93.2	91.4	0.058	0.044	0.326	0.226	$5.6\times$	$5.2\times$
`diabetes_reg`	10	2.76	3.81	94.5	92.0	0.067	0.051	0.409	0.299	$6.1\times$	$5.9\times$
`heart`	13	8.26	9.72	98.3	93.3	0.014	**0.070**	0.324	0.324	$23.1\times$	$\mathbf{4.6}\times$

The claim survives; one headline number is capacity-specific. The minimal explanation still spends almost all of the budget (85.9–93.3% against the 87.4–98.3% of the arm this one differs from by a single change), the certified radius still rises monotonically as ε_{sel} tightens on all four datasets, and tightening still buys radius for a fraction of a feature — so the mechanism, the fix and the shape of the frontier are properties of the criterion and not of the small net. What does *not* transfer is the size of the gain we quote first. On `heart` it is $23.1\times$ at $(128, 128)$ and $\mathbf{4.6}\times$ at the canonical capacity, and the reason is that the canonical model's *minimal* explanation starts from a five times larger radius ($0.014 \to 0.070$), not that the stabilised one got worse (0.324 at both capacities). Across the four datasets the gain goes from 2.9–$23.1\times$ to 2.4–$5.9\times$. Neither is the headline: the body quotes App. A48, which repairs this defect and the validation split together, and the $(128, 128)$ arm remains only where its companion runners still use it (App. A44 lists them).

Why radii are mostly smaller at higher capacity here — and why that does not survive App. A48. A wider encoder has larger operator norms, so the certified analytic Lipschitz constant — and with

it the padding $L_i h$ — grows, and on three of four datasets the stabilised radius is indeed smaller. But the cell that moves most, `heart`'s *minimal* radius ($0.014 \to 0.070$), moves the other way, and App. A48 shows why: with a genuine validation split it is 0.009 at this capacity and 0.008 at the small one. The large in-sample value is a product of early-stopping a wider model on rows it trained on, not of capacity. Artifact: `results/explanation_stability/canonical_frontier_G1e6_seed42.json` and its seed 43–46 companions.

A48 The Frontier Under a Protocol With No Known Defect

Bound: unsigned (arm A) throughout this appendix, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: canonical capacity and a held-out validation split — both repairs at once, the protocol the body quotes.

Two appendices repair one disclosed defect each and leave the other in place. App. A47 rebuilds the frontier at the accuracy benchmark's own capacity, but validates on `Xtr[:200]`; App. A44 partitions the training rows $80/20$ before the model is built, but at the small $(128, 128)$ encoder. Both bind hardest on `heart`, which is the dataset the headline rests on, and in opposite directions — so neither appendix could say what the headline is worth once both defects are gone. `run_stability_frontier.py --canonical --holdout_val 0.2` composes them. Five seeds, $G{=}10^6$, same evaluation points.

Table 40: The frontier with both disclosed defects repaired at once: benchmark capacity *and* a genuine validation split. Five seeds (42–46), 40 instances each, means over seeds. $\varepsilon{=}0.5$; $\varepsilon_{\text{sel}}{=}0.35$ in the right-hand columns. The last column is the share of instances pinned at $r_{\max}{=}0.5$: the *gain* column is therefore a ratio of a 6–49%-censored mean to a nearly uncensored one on two datasets; it is a lower bound only where the denominator has no capped instance. `heart`, which carries the headline, is 0% censored at minimality and 19% at 0.35.

Dataset (d)	budget used	$\lvert S\rvert$	certified radius	gain	certifiable	censored
`abalone` (8)	85.2%	3.21→3.79	0.154 → 0.387	2.5×	98→100%	6→49%
`wine` (11)	90.5%	4.15→5.29	0.050 → 0.243	4.9×	94→100%	0→6%
`diabetes_reg` (10)	93.7%	3.10→4.10	0.073 → 0.384	5.3×	96→100%	1→39%
`heart` (13)	**98.7%**	7.93→8.76	**0.009 → 0.286**	**31.7×**	90→100%	0→19%

The claim survives, and the two repairs do not compound. Every qualitative statement holds: the minimal explanation still spends 85–98% of ε at the unperturbed point, the radius it leaves is still near zero on `heart`, and 0.8 extra features still buy $31.7\times$. Quantitatively the two single-defect controls predict different things and neither predicts the composed cell. Read alone, capacity *shrinks* `heart`'s gain ($23.1\times \to 4.6\times$, App. A47) and the clean split *grows* it ($23.1\times \to 29.7\times$, App. A44); composed, it is $31.7\times$. The resolution is that the capacity effect was never a capacity effect: a wider model early-stopped on its own training rows starts from a minimal radius of 0.070, and once validation is genuine that radius is 0.009 at the benchmark capacity and 0.008 at the small one. Reading either control alone would have mispredicted the composed result, in opposite directions, which is the argument for running it rather than reasoning about it.

It is not one seed. Per-seed `heart` gains are 35.9, 71.0, 24.2, 23.1 and $26.9\times$. We quote the ratio of the seed-mean radii ($31.7\times$); the per-seed mean is $36.2\times$, the geometric mean $32.9\times$ and the median $26.9\times$. A ratio whose denominator is near zero is unstable under averaging, so the four summaries are all given; none is below $23\times$.

What this does and does not settle. It settles the composition question for the *frontier*. It does not rebuild the runners that still share the `Xtr[:200]` split (App. A44 lists them): the budget-consumption and frontier tables of App. A54, the set-stability comparison, the MIP-fragility and certificate-1 controls, the EBM frontier and the d-sweep. The across-backbone headroom, the form-fidelity control and the formal-XAI comparison have since been re-run on this protocol. Artifacts: `results/explanation_stability/clean_frontier_G1e6_seed42.json` and its seed 43–46 companions.

A49 CERTIFYING AGAINST THE DISPLAYED RESIDUAL, NOT THE SURROGATE

This appendix holds the paper's central measurement: the frontier's $31.7\times$ on `heart` under the explainer's own unsigned acceptance test is $1.24\times$ once the same explanation is certified against the predicate it actually makes — a factor of 25 between the two bounds, and the reason the paper's headline is the bound rather than the knob.

Why selection may use magnitudes and certification may not. The two steps answer different questions, so they may use different predicates without contradiction. Selection answers *which features does the explanation name*; a user reads an omitted feature as "did not matter here", and two large contributions that cancel at x did matter, so the per-feature magnitude test is the right relevance criterion — selecting on the signed sum (arm C) returns explanations that omit such pairs, down to the empty explanation "the intercept is within ε", a true but much weaker statement. Certification answers *is the statement displayed true across the ball*; that statement is $|f(x') - g(x')| \le \varepsilon$, and whether it is true depends on the signed sum, cancellation included. Charging it the unsigned sum certifies a stronger predicate than the one shown, which is sound but is not a certificate of the explanation. A reader who wants the omitted features individually small everywhere on the ball wants arm A's predicate, and should use it, knowing it certifies little near minimality.

What a ratio here estimates. Every mean radius in this paper is a mean of $\min(r, r_{\max})$, the radius the bisection returns when its search is capped at $r_{\max}$. Per arm that mean is a lower bound on the uncapped mean. A *ratio* of two such means is not a bound in either direction once both arms have instances at the cap: paired radii $(1, 100)$ and $(2, 100)$ have ratio of means $102/101 = 1.01$, and capped at 10 the ratio reads $12/11 = 1.09$ — larger. A capped ratio is a lower bound on the uncapped one only when its denominator has no instance at the cap. We therefore report ratios as ratios of capped means, name the cap, and lean on two estimands truncation cannot distort: the share of instances certifying $r \ge t$ for a threshold t below the cap, and paired differences at a stated cap. Where an earlier revision called a ratio with a censored denominator a "lower bound", the sentence is corrected below.

Raising the cap and raising $r_{\max}$ are different experiments. `--r_max` sets both the cap and the error table's margin, so a run at $r_{\max}$=2.0 certifies different forms (and sometimes different subsets) from a run at 0.5. Holding the $r_{\max}$=2.0 explanations fixed and capping only the search at 0.5 isolates the cap: on `heart` the bound's ratio goes from $\mathbf{19.1}\times$ uncapped to $\mathbf{14.7}\times$ capped (arm A is never at the cap, so this is a lower bound), and the knob from $1.48\times$ to $1.23\times$; over the four datasets the capped ratios are $2.3/5.4/6.2/14.7\times$ against $3.2/6.0/10.5/19.1\times$ uncapped. The rest of the distance to the $37.8\times$ quoted at $r_{\max}$=0.5 is the table, not the cap: the narrower table admits cheaper forms, which spend more of the unsigned budget, and arm A's minimal radius falls from 0.024 to 0.009. By the threshold estimand the conclusion does not depend on either choice: at $r_{\max}$=2.0, arm A certifies $r \ge 0.1$ on $53/14/24/2\%$ of minimal explanations (`abalone`/`wine`/`diabetes_reg`/`heart`) and arm D on $93/94/99/93\%$; at $r \ge 0.25$ the shares are $28/2/3/0\%$ against $72/57/89/74\%$. Computed from the per-instance radii in `armd_uncensored_seed{42..46}.json` (`scripts/fixed_explanation_cap.py`).

The mismatch. Section 4 says of the interval decomposition that "the *signed* form matters: the naive $\sum_i \max(U_i, -L_i)$ ignores cancellation and can overstate the worst case." Section 3.2 then defines VIB-X's acceptance test with the unsigned $\sum_{i \notin S} |f_i(x_i)|$, and says so explicitly: it "upper-bounds the displayed residual" and sign cancellation "could admit cheaper explanations." Using a conservative test to *select* is a defensible choice — it is a per-feature relevance test, and an explanation should not omit a feature merely because two large contributions happen to cancel at this x. Certification inherited the same unsigned form, which is the defect: the radii an unsigned criterion certifies do not certify the predicate the explanation states. The displayed surrogate's error at a perturbed point is

$$f(x') - g(x') \;=\; \underbrace{\textstyle\sum_{i \notin S} f_i(x'_i)}_{\text{signed}} \;+\; \textstyle\sum_{i \in S} (f_i - g_i)(x'_i),$$

whose worst case over the ball is bounded by $\max\big(\sum_{i \notin S} U'_i,\, -\sum_{i \notin S} L'_i\big) + \sum_{i \in S} e_i^{\mathcal{C}_i}$ with $[L'_i, U'_i]$ the padded interval of f_i itself. That bound is sound, never looser, and strictly tighter whenever the per-

feature $|\cdot|$-maxima are not all attained in a common direction. That condition is weaker than "excluded contributions disagree in sign" and fires more often: two features whose contributions are both positive at x can still tighten the bound $6.0 \to 5.0$. (`tests/test_signed_criterion.py` brute-forces soundness: no violation over 250 random additive models with an exact error table ($e_i \equiv 0$), 200 more carrying a residual table, and 80 one-dimensional shapes on which the signed range is checked to bracket the function over the ball. Be clear about what that does and does not establish: the brute-force reference computes its own worst case by the same coordinatewise separation the proposition asserts, so these tests exercise the padding and the error-table handling rather than separability itself. Separability is what Prop. 4 proves, and it is proved rather than sampled.)

Three arms. All on the same models and instances, canonical capacity with a held-out validation split, five seeds. **A** selects and certifies unsigned — the explainer's own acceptance certificate, and the frontier the body reports beside arm D. **B** keeps A's selections and certifies signed — the requested experiment, and the one that certifies the stated predicate. **C** uses the signed quantity for selection as well.

Why these are threshold statistics. Under B the radii run into the censoring ceiling $r_{\max}$=0.5 on 12–53% of instances at the minimal explanation and 35–80% at ε_{sel}=0.35, so a mean radius there is not a measurement. The fraction certifiable at any threshold *below* the ceiling is exact, because censoring only truncates from above. Raising $r_{\max}$ does not fix this cleanly: the error table must cover the ball, so a wider margin inflates every $e_i^{\mathcal{C}}$ and changes which explanations the greedy returns.

So the two $r_{\max}$ runs are two measurements, not one measurement at two ceilings. On `heart` at minimality neither run is censored at all, and the arm-A mean still moves $0.009 \to 0.024$: that is entirely the margin. The arithmetic settles the direction — decensoring the ε_{sel}=0.35 column alone would take the knob ratio to 38.9×, *above* the censored run's 31.7×, where the $r_{\max}$=2.0 run reads 14.9×. In the MIP-fragility study the same substitution moves the arm-A radius on 99 of 100 instances while only 1 ever sat at the 0.5 ceiling, with $|S|$ identical on all 100: the selections hold and the error table moves. We therefore report both runs and label each, and we claim no frontier ratio is invariant to $r_{\max}$.

Table 41: Fraction of instances certifiable at radius $\geq t$, five seeds. ε=0.5. Censoring-free: every t is below the ceiling.

	minimal expl. (ε_{sel}=ε), t=0.1			t=0.25, minimal → ε_{sel}=0.35		
Dataset	A	B	C	A	B	C
`abalone`	0.51	**0.91**	0.40	0.24→0.77	0.66→0.84	0.05→0.29
`wine`	0.12	**0.93**	0.46	0.01→0.39	0.54→0.83	0.05→0.12
`diabetes_reg`	0.24	**0.93**	0.52	0.05→0.82	0.85→0.97	0.17→0.49
`heart`	0.00	**0.81**	0.12	0.00→0.48	0.56→0.84	0.01→0.03

Table 42: Arm B's mean radii and its own frontier ratio, five seeds, same models as Table 41. Censored at $r_{\max}$=0.5; arm B's own uncensored $r_{\max}$=2.0 replicate is in parentheses. Every uncensored ratio is larger, as censoring of the 0.35 column predicts, but their *order* across datasets differs from the censored one (1.45–1.72 against 1.16–1.41), so the censored column ranks datasets and the uncensored one measures the effect.

	arm A (paper)		arm B		frontier ratio	
Dataset	minimal	0.35	minimal	0.35	A	B (uncens.)
`abalone`	0.154	0.387	0.348	0.436	2.5×	1.25× (1.72)
`wine`	0.050	0.243	0.283	0.380	4.9×	1.34× (1.51)
`diabetes_reg`	0.073	0.384	0.407	0.474	5.3×	1.16× (1.45)
`heart`	0.009	0.286	0.282	0.397	**31.7**×	**1.41**× (1.70)

The comparison that matters. Table 42 puts arm B's two operating points side by side. It says two things. First, **arm B at the *minimal* explanation matches or beats arm A at the stabilised one —**

`wine` 0.283 against 0.243, `diabetes_reg` 0.407 against 0.384, `heart` 0.282 against 0.286 — at 0.6–1.1 *fewer* features. Second, the frontier gain that survives under arm B is 1.16–$1.41\times$ censored and 1.45–$1.72\times$ uncensored, against the 2.5–$31.7\times$ of §5.3. So on our own measurements the free change (certify with the tighter bound) is worth more than the change we sell (buy extra features), and a reader deciding what to do should take the bound first.

The mechanism as narrated does not survive at its stated magnitude either. The budget-consumption figure quoted throughout this paper — 85.2–98.7% of ε spent at the unperturbed point, "leaving nothing for a perturbation to occupy" — is arm A's. Applying the same definition to the signed slack this runner already records gives $46.7/36.1/41.5/53.3\%$. Under the predicate we now call correct the greedy spends *about half* its budget, which is why 81–94% of arm-B minimal explanations certify $r \geq 0.1$. What survives is narrower and should be stated as such: the greedy exhausts the budget of *its own acceptance test*, and that test is not the quantity that governs robustness.

What this costs the paper. The sentence "the radius is near-zero for the minimal explanation" is true of arm A and not of arm B. Certified against the residual it actually displays, the same minimal explanation on `heart` certifies $r \geq 0.1$ on 81% of instances rather than 0%, and its mean radius rises from 0.009 to 0.28 — a geometric-mean factor of 36.3 on `heart` and 3.7–10.9 elsewhere. The frontier survives as a direction — more selection budget still buys radius, `heart` $0.56 \to 0.84$ at t=0.25 — but it is a far smaller effect than the $31.7\times$ of the unsigned arm, and a reader should treat that figure as a property of the conservative surrogate criterion rather than of minimality as such. The body quotes both arms, labelled (§5.3).

Arm D: signing the other half. Arm B signs the excluded features but still charges each included one $\pm e_i^{\mathcal{C}_i}$, a non-negative supremum over the error table's *whole* tabulated range. That discards a sign — a residual that only ever runs positive cannot cancel against an excluded feature that only ever runs negative, though $\pm e_i$ says it can — and a locality: the supremum spans several SD of a standardised feature while certification needs it only on a ball of radius $r \leq 0.5$. Signing and localising the included terms removes a third of arm B's consumed budget on `heart` ($53.3\% \to 35.5\%$ of ε), so it is worth measuring. `ErrorTable.signed_residual_range` answers it, using the same superset window and the same $L_i^{\mathrm{res}} h/2$ padding the global constant carries, so it is the same bound on a smaller set kept signed. The resulting criterion,

$$\max\Big(\textstyle\sum_{i \notin S} U_i' + \sum_{i \in S} U_i^e,\ -\big(\sum_{i \notin S} L_i' + \sum_{i \in S} L_i^e \big) \Big),$$

is never looser than arm B — and it is not merely the tightest bound we tried. It is *exact*, up to the certified grid padding.

Proposition 4 (Arm D is the worst case, not a bound on it, for a main-effects model)**.** *Fix x, r and a displayed surrogate $(S, \mathcal{C})$ of a model whose shape functions are univariate. Over $\|\delta\|_\infty \leq r$ the ball is a product set, and the surrogate's error $\sum_{i \notin S} f_i(x_i{+}\delta_i) + \sum_{i \in S}(f_i{-}g_i)(x_i{+}\delta_i)$ is a sum of one-dimensional functions of* distinct *coordinates. Each term may therefore be maximised (or minimised) independently, so $\max_\delta \sum_i t_i(\delta_i) = \sum_i \max_{\delta_i} t_i(\delta_i)$ and the two extremes are* attained*. Hence $\max_\delta |\cdot| = \max\big(\sum_i U_i, -\sum_i L_i \big)$, which is arm D's criterion. Its only conservatism is the same $L_i^{\mathrm{res}} h/2$ padding and one-knot superset window the certificate carries everywhere (§4), so arm D is the exact worst case up to $O(Lh)$.*

Attainment is a property of the main-effects model, not of arm D. The premise is that the error decomposes over *distinct* coordinates. On an interaction-bearing model it does not: the displayed surrogate carries no two-dimensional term, so every h_{ij} is excluded and enters the same sum, and f_i and each h_{ij} containing i share x_i. Maximising them separately is then a relaxation — sound, not attained — exactly as for the model-level interval in App. A23. Every radius the body quotes for arm D is measured on a main-effects model, where the premise holds; the one exception is App. A24, which certifies the interaction-bearing model and whose arm-D radii are therefore sound but not attained.

This is the same separability argument Eq. 2 already makes for the robustness certificate: it turns "we tried tighter bounds and they stopped moving" into "no sound bound over this ball can do better, up to the $O(Lh)$ grid padding and the outward-rounded superset window", which is the claim that actually matters. The empirical regress below is then a consistency check on the proposition rather than the argument for it.

Table 43: All three certification bounds on the same greedy selections, five seeds, canonical capacity and a clean split. Every column is sound; they differ only in how much cancellation they decline to throw away. **Arm D is censored harder than arm B**: $53/15/62/30\%$ of instances sit at $r_{\max}{=}0.5$ at minimality and $75/36/84/52\%$ at $\varepsilon_{\text{sel}}{=}0.35$, against arm B's 12–$53\%$ and 35–80%. The 0.35 column is censored harder than the minimal one; cap shares alone do not fix the direction of a ratio of capped means (App. A49), and in the fixed-explanation control there the capped ratio came out smaller. Re-run at $r_{\max}{=}2.0$ the ratio is $\mathbf{1.40}$–$\mathbf{1.58}\times$ rather than 1.11–$1.30\times$; censoring there is 0–4% at minimality but 0.5–14% in the 0.35 column that forms the numerator, so on `abalone` and `diabetes_reg` these are ratios of capped means with both arms capped, not bounds (App. A49) (`heart`, censored on 3% there, is $1.48\times$ at a minimal radius of 0.454). The wider ceiling leaves the *subsets* alone and not the forms: the two runs give *bit-identical* $|S|$ on all four datasets at every budget, so no feature enters or leaves, but a wider error-table margin makes simple forms infeasible more often, so the same features are displayed more faithfully — which is why `heart`'s arm-A mean radius reads 0.024 here against 0.009 at $r_{\max}{=}0.5$, the replicate two paragraphs below.

	mean radius, minimal expl.			share with $r \geq 0.1$		frontier
Dataset	A	B	D	A	D	ratio, D [95%]
`abalone`	0.154	0.348	**0.375**	0.51	0.94	$1.17\times$ $[1.08, 1.26]$
`wine`	0.050	0.283	**0.293**	0.12	0.94	$1.30\times$ $[1.12, 1.54]$
`diabetes_reg`	0.073	0.407	**0.434**	0.24	0.99	$1.11\times$ $[0.98, 1.29]$
`heart`	0.009	0.282	**0.342**	0.00	0.92	$1.24\times$ $[1.06, 1.48]$

The regress terminates at D, and the answer does not change. Arm D buys 1.04–$1.21\times$ on top of arm B — most of it on `heart`, which is where the included terms were largest. **Dispersion, and what it licenses.** The frontier ratio's between-seed 95% t-interval (on the log ratio, five seeds) is in the table's last column: censored, it excludes 1 on three of four datasets and *includes* it on `diabetes_reg` ($[0.98, 1.29]$), so the censored knob is not distinguishable from nothing there. Off the ceiling every interval excludes 1 ($[1.48, 1.68]$, $[1.22, 1.82]$, $[1.18, 1.77]$, $[1.35, 1.65]$), which is the form the claim should be read in: the knob is worth 1.4–$1.6\times$ and that is bounded away from 1; the censored 1.1–$1.3\times$ is a ratio of means capped on both sides, which is neither a measurement of the uncapped ratio nor a bound on it. It does not reopen the question: the budget consumed falls only from 36–53% to 32–38%, the share certifying $r \geq 0.1$ from 81–94% to 92–99%, and the uncensored frontier ratio from 1.45–$1.72\times$ to 1.40–$1.58\times$. So the conclusion is not that every tightening keeps shrinking the effect indefinitely; it is that essentially all of the shrinkage happens at the first step, from the explainer's unsigned acceptance test to the signed residual, and that step is free. The body quotes both, arm D as the bound that governs the displayed explanation and arm A as the explainer's own acceptance certificate.

Is ε even comparable across datasets? No — and correcting it makes the effect larger. App. A46 calibrates ε across *backbones* and argues that an uncalibrated budget could manufacture an entire spread. The same argument applies one level up, across *datasets*: the three regression sets have a standardised target, so $f(x)$ lives on a scale of order one, while `heart` is a binary head whose logits are standardised by nothing. Measured over five seeds at canonical capacity, σ_{pred} is $0.78/0.53/0.76$ on `abalone`/`wine`/`diabetes_reg` and $\mathbf{2.24}$ on `heart` — so at a shared $\varepsilon{=}0.5$ the dataset carrying every bolded number in this paper was certified against a relative budget 2.9–$4.2\times$ tighter than the others, for a reason with nothing to do with minimality — a defect in the comparison, not a property of the explanations.

The calibrated arm (`armd_calib`, written by the same runner with `--calibrate_eps`) prices ε and every ε_{sel} as the same fraction of each dataset's own σ_{pred}, giving `heart` an effective ε of 1.12 rather than 0.5.

The confound is real and it was running against us. The obvious hypothesis is that an uncalibrated ε manufactures `heart`'s fragility. It is not: correcting it moves `heart`'s arm-A ratio from $31.7\times$ to $35.6\times$, so the shared-ε number we have been reporting is the *conservative* one. The mechanism explains the direction. A looser budget lets the greedy drop *more* features, so the minimal explanation gets smaller and therefore more fragile — which is this paper's own claim, sharpened rather than

Table 44: ε priced per dataset against its own prediction spread. Five seeds, canonical capacity, clean split. The two arms are *two runs* of the same runner (`armd` and `armd_calib`) differing only in `--calibrate_eps`; they train the same models, which is checkable in the artifacts — σ_{pred} agrees to every recorded digit on all four datasets. The frontier ratio is the minimal explanation against ε_{sel}=0.35 in absolute output units, i.e. 0.7ε — the same operating point as everywhere else in the paper, and in the calibrated arm both ε and ε_{sel} are scaled by the same σ_{pred}, so the ratio of the two is unchanged. The arm-D ratios are ratios of capped means: 11–53% of instances sit at $r_{\max}$=0.5 at minimality and 26–76% in the 0.35 column. Cap shares alone do not fix the direction of the distortion (App. A49); for the fixed-explanation control there, the capped ratios came out smaller.

			arm A ratio		arm D ratio	
Dataset	σ_{pred}	ε_{eff}	shared	calibrated	shared	calibrated
`abalone`	0.78	0.39	2.5×	2.7×	1.17×	1.15×
`wine`	0.53	0.26	4.9×	3.6×	1.30×	1.33×
`diabetes_reg`	0.76	0.38	5.3×	6.5×	1.11×	1.12×
`heart`	**2.24**	1.12	31.7×	**35.6**×	1.24×	1.21×

dissolved. This is the one control here that runs in our favour, so we state its limits plainly: it is one control on four datasets, `wine` moves the opposite way (4.9× → 3.6×), and we continue to quote the shared-ε numbers in the body because they are lower. What matters for the body's claim is that arm D — the bound it now quotes — is nearly unmoved: 1.11–1.30× shared against 1.12–1.33× calibrated. Artifacts: `results/explanation_stability/armd_calib_seed42.json` and its seed 43–46 companions.

Selecting on the signed quantity is not the repair. Arm C exhausts a smaller budget (59–81% against A's 85–99%) and its radii sit between A and B on three datasets (below A on `abalone`), so the exhaustion mechanism itself is robust to the change. But the explanations degenerate: $|S|$ falls from 3.21/4.15/3.10/7.93 to 0.94/0.52/0.78/3.17, and the returned explanation is *empty* on 52%, 61%, 45% and 14% of instances. Under the predicate an empty explanation asserts only that the intercept approximates f within ε here, which can be true; what it cannot convey is that two large contributions cancel at this x, which a reader of "no feature named" will take as "none mattered". That is the argument for keeping an unsigned *selection* rule, and it is why B — unsigned selection, signed certification — is the combination we would now call correct.

The ceiling is not doing the work. Repeating the whole comparison at $r_{\max}$=2.0 removes the censoring (0–3% of instances at the ceiling under B, against 12–53% before) and reproduces the result: at t=0.1 arm B certifies 0.91/0.93/0.94/0.82 against 0.53/0.14/0.24/0.03 for arm A. This is a robustness check rather than a cleaner version of the same measurement — a wider $r_{\max}$ forces a wider error-table margin, which inflates every $e_i^{\mathcal{C}}$ and changes which *forms* the greedy can afford (the subsets do not move: $|S|$ is bit-identical on all four datasets at every budget); on `abalone` at ε_{sel}=0.05 it makes every simple form infeasible and the explainer degenerates to plain feature selection, which the soundness warning catches. The two runs agree to within two points at t=0.1 on every dataset; the one visible difference is `heart`'s arm-A mean radius (0.009 against 0.024), for that reason. The gap between A and B is what is stable.

Artifacts: `results/explanation_stability/armd_seed42.json` and `results/explanation_stability/armd_uncensored_seed42.json` with their seed 43–46 companions (`elasticnam/experiments/run_signed_criterion.py`); arms A, B, C and D are recorded in the same run on the same models.

A50 The Additive Gauge: Does Centring the Shapes Change the Headline?

An additive model is identified only up to constants: for any c with $\sum_i c_i = 0$, the shapes $f_i + c_i$ predict exactly what f_i do. The explainer's acceptance test $\sum_{i \notin S} |f_i(x_i)|$ and its ranking by $|f_i(x_i)|$ are *not* invariant to that choice: adding $+c$ and $-c$ to two excluded shapes leaves every prediction and the signed residual unchanged and inflates the unsigned sum without limit. Arm D is not invariant

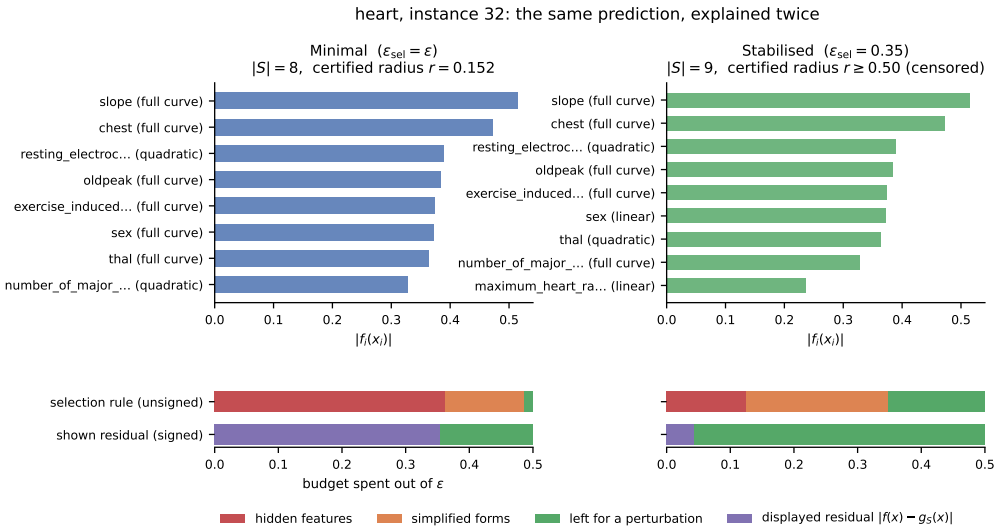

Figure 12: One reader's view: one `heart` prediction explained twice (clean protocol, arm D). One extra feature and two simpler forms move the certified radius from 0.152 to the ceiling. The bars say why twice over: the unsigned rule has spent 0.487 of ε=0.5 on the minimal explanation while the signed residual the reader sees is 0.354 — that gap is the bound this paper is about — and the stabilised explanation leaves the signed budget almost untouched (0.043). **The largest-gap instance of** 40**, on seed** 42: on that seed the median `heart` minimal explanation already certifies 0.472 under arm D and 42% of them sit at the ceiling (pooled over five seeds, 0.354 and 30%), so this card is the tail where the knob still buys something. `paper/figures/make_stability_card_fig.py`.

either, in a milder way: the displayed surrogate's intercept absorbs whatever constants the excluded shapes carry, so the predicate it certifies, $|\sum_{i \notin S} f_i(x'_i) + \cdots| \leq \varepsilon$, is measured from a different zero. VIB-NAM imposes no gauge — each shape carries whatever offset training left it — so the headline ratio could in principle be a property of how one network split its constants rather than of the model's function.

The control. We fix the conventional gauge — every shape centred to mean zero over the training rows, the offsets moved into the intercept, as EBM's fitted terms and the exact additive Shapley value already are — and recompute arms A and D on the *same* trained models, splits and instances, paired against the raw gauge. Prediction is unchanged (asserted per instance); the certifier, the error table and the greedy all read the shapes through one accessor, so the gauge is the only thing that moves. Canonical capacity, held-out validation split, five seeds, G=10^6, $r_{\max}$=0.5 — the protocol of the body's headline, whose numbers the raw column reproduces exactly.

Table 45: Raw / centred gauge on the same models (five-seed means). "same S" is the share of instances whose minimal subset is identical under both gauges. Radii at ε_{sel}=ε (minimal explanation); the knob is the ε_{sel}=0.35 arm-D radius over the minimal one; the last column is the share of minimal explanations arm D certifies at $r \geq 0.1$. Ratios are of capped means (App. A49); arm A is at the cap on $\leq 6\%$ of instances and arm D on 6–62%.

Dataset	$\lvert S\rvert$	same S	arm A r	arm D r	D/A	knob	D: $r \geq 0.1$
`abalone`	3.21 / 2.86	52%	0.1536 / 0.1565	0.3752 / 0.3631	2.4 / 2.3	1.17 / 1.21	94 / 88%
`wine`	4.15 / 3.50	13%	0.0497 / 0.0590	0.2929 / 0.2783	5.9 / 4.7	1.30 / 1.32	94 / 96%
`diabetes_reg`	3.10 / 2.26	18%	0.0729 / 0.0868	0.4335 / 0.3915	5.9 / 4.5	1.11 / 1.21	99 / 96%
`heart`	7.93 / 7.75	9%	0.0090 / 0.0095	0.3423 / 0.3133	37.8 / 33.1	1.24 / 1.31	92 / 91%

What the gauge changes, and what it does not. It changes the explanation: on `heart` the centred minimal subset is identical to the raw one on only 9% of instances, and centring shrinks $|S|$ on every dataset. It does not change the finding: the bound is worth 2.3–$33.1\times$ centred against 2.4–$37.8\times$ raw (geometric mean $6.4\times$ against $7.5\times$), the knob 1.21–$1.32\times$ against 1.11–$1.30\times$, and by the threshold estimand, which truncation cannot distort, arm D certifies $r \geq 0.1$ on 88–96% of centred minimal explanations. The raw gauge flatters the headline by at most $1.3\times$ (`diabetes_reg`); it does not manufacture it. We keep the raw gauge in the body because every other artifact uses it, and state the centred figures beside it. Artifacts: `results/centering_control/centering_seed42.json` and seeds 43–46 (`elasticnam/experiments/run_centering_control.py`, `scripts/centering_summary.py`).

A51 `BIKE` IS AN HOURLY SERIES AND WE SPLIT IT RANDOMLY

`bike` is OpenML 42712, the UCI Bike Sharing *hourly* file: one row per hour over two years, with `year`, `month`, `hour` and `weekday` among its twelve features. Its target has a lag-1 autocorrelation of $+\mathbf{0.844}$ and $+0.816$ at lag 24. Every split in this paper uses `KFold(shuffle=True)`, so adjacent hours land on both sides of every fold boundary, so `bike`'s scores are optimistic for a time series and we measure by how much below.

`run_full_benchmark.py --temporal_split` replaces the shuffle with contiguous blocks in row order, which for this file is time order.

Table 46: `bike` R^2 under the shuffled split Table 24 reports and under a contiguous temporal split. Five folds, seed 42, same nested protocol otherwise; the temporal column carries its fold SD and the last column is each method's mean per-fold rank (average ties). The temporal arm ran on CPU and the shuffled one on a GPU, so Δ carries device noise as well, which we did not measure separately. Contiguous folds are not a forecasting protocol: four of the five train on later hours than they test on, so this measures the leakage of adjacent hours, not forecast skill.

Method	shuffled	temporal	Δ	rank of means	mean fold rank
XGBoost	0.952	0.802 ± 0.149	-0.150	$1 \to \mathbf{4}$	2.8
EBM	0.923	0.816 ± 0.089	-0.107	$2 \to 2$	3.0
NODE-GAM	0.897	0.816 ± 0.070	$\mathbf{-0.081}$	$3 \to \mathbf{1}$	3.4
VIB-NAM-Boost-Ens	0.895	0.811 ± 0.080	-0.084	$4 \to \mathbf{3}$	2.8
VIB-NAM-Boost	0.894	0.801 ± 0.084	-0.093	$5 \to 5$	3.8
VIB-NAM	0.841	0.751 ± 0.087	-0.091	$6 \to 6$	5.2
NAM	0.693	0.543 ± 0.155	-0.150	$7 \to 7$	7.0
Linear	0.394	0.271 ± 0.054	-0.123	$8 \to 8$	8.0

The leakage is real, it inflates everything, and `bike`'s ranking does *not* survive it. Every arm loses 0.08–0.15 R^2, so the shuffled `bike` column in Table 24 is optimistic in absolute terms. **Measured with NODE-GAM excluded and the inner validation split still random, the ordering reads $7/7$, and that measurement is wrong**: NODE-GAM defaults off in the runner, and a random inner split leaves early stopping choosing against a signal in which 96% of validation rows have an adjacent hour in training. With NODE-GAM included and the inner split made contiguous, the ordering of means is $5/8$: XGBoost falls from first to fourth and NODE-GAM rises from third to first. The two arms whose ranks move are precisely the ones the leaked validation signal helped most — XGBoost tunes 300 rounds against it and NODE-GAM early-stops on it — so excluding NODE-GAM removed the arm most likely to move, and the random inner split hid the movement. **The order among the top four is not resolvable at $n{=}5$ folds**: they sit within 0.014 R^2 of one another under fold SDs of 0.07–0.15, their mean per-fold ranks are 2.8–3.4 with XGBoost tied first per fold, and the paired per-fold differences between adjacent pairs are $+0.000$ $[-0.041, +0.041]$ (NODE-GAM$-$EBM), $+0.005$ $[-0.018, +0.027]$ and $+0.009$ $[-0.077, +0.096]$, each winning $2/5$ folds. What the folds support is that the shuffled ordering does not survive, not a new one.

What the aggregate claims do, and do not, survive. §5.1's inference is on ranks over nine datasets, not on absolute scores, so the question is what `bike`'s new column does to those. Recomputed

with bike's temporal row substituted: average ranks move on three of eight methods and by at most 0.33 (XGBoost $4.22 \to 4.56$, NODE-GAM $4.89 \to 4.67$, ours $3.28 \to 3.17$; average ties, the convention of Table 24), the average-rank *ordering* is unchanged, EBM remains best at 2.89, and Friedman moves from p=0.058 to p=0.055 — above α either way. So the aggregate conclusions hold and bike's own ordering does not (results/benchmark/temporal_ranks.json, scripts/temporal_ranks.py).

What we have not done. The temporal split is bike alone at seed 42; Table 24 still reports the shuffled column and the Holm family is not re-run. Four of the five folds train on the future — contiguous K-fold is not forward chaining, and TimeSeriesSplit is the instrument that would be — so the residual leakage through year, month and the series trend is untouched. bike is also the leverage point in the fit control (App. A39) and the largest main-effects-only certification cost, and neither is re-measured here. Artifact: results/benchmark/bike_temporal.json (run_full_benchmark.py --datasets bike --temporal_split).

A52 Two Model Populations Under One Protocol Label

This appendix exists because we shipped a fix without the run that makes it true. run_stability_frontier.py seeded once before its dataset loop, so a dataset's model depended on how much randomness the datasets before it had consumed and --datasets heart did not reproduce the heart column of a four-dataset run. We changed it to seed per (dataset, seed) — and did not regenerate the artifacts the old code had already written. For one revision the paper therefore carried two model populations under a single protocol label.

The fix reached one runner, not the class. A pre-submission audit checked the others and found the same shape: run_full_benchmark.py (Table 24), run_exact_certification.py and run_ablation.py still call set_seed once before their dataset loop, and run_matched_baselines.py does not call it at all. Each of those artifacts is therefore reproducible by *its own command*, which is how every one of them was produced and what REPRODUCE.md records — but a single-dataset invocation will not reproduce that dataset's column, exactly as above. We are disclosing rather than repairing this three days before the deadline for the reason the rest of this appendix gives: changing the seeding now would make the shipped numbers unreproducible by the shipped code, which is a worse failure than an order-dependent stream. The frontier runner, which carries the headline, is the one that was fixed.

How it showed. Two independent signatures. *Order.* stability_frontier_G1e6 and the headroom runner share a configuration exactly; in the affected revision, on abalone, the *first* dataset in the list, they agreed bit-for-bit on all five seeds, and on wine, diabetes_reg and heart they diverged completely. That is what seeding once before the loop looks like from outside (they now agree on heart to the third decimal). *Size.* At $\varepsilon_{\text{sel}}=\varepsilon$ the pre-fix artifacts give mean $|S|$ of $3.21/4.25/2.89/8.35$ and the post-fix ones $3.21/4.15/3.10/7.93$. Two tables drawing on the two populations is what the gate is built to refuse: an artifact whose producing runner has changed behaviour since it ran now fails the build, and a family whose seeds were written by two behaviours fails it too.

What we did. Every cited five-seed family written by an affected runner was regenerated under the shipped seeding — the frontier in all four protocol arms (Tables 40, 39, 35 and App. A54), the four certification arms and their $r_{\max}$=2 and ε-calibrated repeats (Tables 41, 42, 43, 44), the identity certificate, the set-stability comparison, the MIP-fragility and certificate-1 controls, the noise injection, the form-fidelity control (whose runner had never seeded the model at all, and now records both bounds on one population), the d-sweep and its L_2 companion, and the formal-XAI cross-evaluation (the last three had the same once-before-the-loop seeding, now fixed). Almost every cited artifact carries a hash of its producing script's behaviour — the nine exceptions are listed two paragraphs down and printed on every build: recorded at run time for the runs made after the field existed, and *backfilled* for 29 artifacts written before it did, each with a reason written into the artifact (scripts/backfill_script_hash.py: a sibling run of the same file that recorded the hash, a bit-identical $|S|$ column across runs, or a reviewed runner diff). A gate step, scripts/check_artifact_freshness.py, compares that hash (or, for older artifacts, the script recovered at

the recorded `git_sha`) with the runner as it stands, checks every sibling seed and not only the one the index names, and *fails* the build if a cited artifact is behind its runner, if the seeds of one family were not all written by the same runner behaviour, or if a cited artifact cannot be checked at all. The comparison is on the *abstract syntax tree with docstrings stripped*, not on file bytes — a byte hash fires on every comment edit, and a gate that condemns every artifact in the repository when a docstring is reworded is a gate that will be ignored within a week. An AST diff fires on a moved `torch.manual_seed` and not on a reworded comment, which is the distinction that matters. The manifest checks that an artifact has not changed; it cannot check that the code which produced it has not.

What remains. Nine cited artifacts record a commit that is not in this repository's history, or were written from a dirty tree *and* carry no behaviour hash — which is the gate's actual rule; 161 artifacts are marked dirty, and the rest are accepted because a behaviour hash pins what the runner computed (the main nested-CV accuracy table, one depth-cell seed, three deep tabular baselines, the worked-examples table, the ablation table and the two greedy-vs-exact optimality arms of Table 23); their runner cannot be recovered and each is a multi-hour rerun, so they are accepted by name with that reason in the gate's source, which prints the list on every build. The depth cell's other seeds are accepted under a named waiver that expires the moment their runner changes again (it has since gained a recorded fit score but computes the same radii), and the two latency studies are exempt from the grid check because their coarser grid is the protocol being timed. Nothing else is excused.

A53 Certifying the Explanation's Identity, Not Just Its Validity

Everything else in this paper certifies that a *displayed* surrogate stays within ε on a ball. That is weaker than what a reader hears in "certified-stable explanation", which is that the explanation would not *change*. We already report the sampled version of the gap and lose it: our re-selection agreement is 0.62 against exact Shapley's 0.71, worse on every seed. The provable counterpart is the object that would settle it, so we built it. Two things about what it certifies. It is the robustness-of-explanations notion of Alvarez-Melis & Jaakkola (2018) made provable rather than sampled — the explanation as a function of the input should not change under a small perturbation of the input — and it is *not* the optimal robust explanation of La Malfa et al. (2021), which asks for a subset that stays a sufficient reason on the ball; here "identity" means, precisely, that the greedy returns the same (subset, forms) everywhere on the ball, i.e. that the algorithm's output is invariant, which is a sufficient condition for the explanation relation to be invariant and not a characterisation of it.

A sufficient condition, and a version of it that was not. VIB-X's output is determined by two things, so we certify both. **(i) Rank stability**: the greedy walks features in descending $|f_i(x_i)|$, so before it can consult the excluded sum at step t the top-t *set* must be pinned — we require $\min_{k<t} \underline{|f_{(k)}|} > \max_{j\geq t} \overline{|f_{(j)}|}$ over the ball, for the prefix the trace actually walks. Ranks at or below the stopping point may permute freely among themselves: they are all excluded, and the excluded sum is determined by the set rather than by any order over it. **(ii) Decision stability**: every $\leq \varepsilon$ comparison the greedy makes must fall the same way everywhere in the ball, i.e. the reachable interval of the compared cost must not straddle ε.

The adjacent-rank form of (i) is unsound, and it is the form one reaches for first. It requires only $\underline{|f_{(k)}|} > \overline{|f_{(k+1)}|}$ between *adjacent* ranks. That pins the relative order *within* the prefix but not the prefix as a set: the reachable intervals are not monotone in rank, so a feature ranked below the prefix can be flat at x and steep nearby and overtake a prefix member it is never compared against. Constants 3.0, 1.0, 0.95 and a line $0.94 + u$ at $\varepsilon{=}2.5$ are a witness — both adjacent tests pass while the fourth feature reaches 1.02 inside the ball and the greedy returns a different subset. The prefix-dominates form above closes it. The numbers in Table 47 are from the prefix-dominates condition; they moved most on `heart` ($0.004 \to 0.002$), and the repair landed together with a fix to per-dataset RNG seeding, so we do not attribute the movement to either alone. Together these make the execution trace identical throughout the ball, so the returned $(S, \mathcal{C})$ is identical. Both are read off the same per-feature intervals the rest of the paper uses; no new machinery.

The condition is sufficient, not necessary, and we verify it the only way that can catch it being wrong: re-running the real explainer at random points of the ball it vouches for (`tests/test_signed_criterion.py`, no disagreement).

Table 47: The ball on which VIB-X provably returns the *same* explanation, against the ball on which the displayed one stays valid (arm D). Five seeds, canonical capacity, clean split, ε=0.5. "rank-bound" is the share of instances where rank stability, not decision stability, is what fails at r=0.05.

Dataset	$\lvert S\rvert$	identity r	certified on	validity r	rank-bound
`abalone`	3.2	0.031	93%	0.375	34%
`wine`	4.2	0.011	68%	0.293	57%
`diabetes_reg`	3.1	0.024	84%	0.434	34%
`heart`	7.9	**0.002**	67%	0.342	**94%**

Identity is certifiable, on a ball one to two orders of magnitude smaller. A positive identity radius exists on 67–93% of instances, so the answer is not "no". But it is **0.005–0.082** of the validity radius: the explanation stays *valid* far past the point where we can prove it stays the *same*. That is the provable counterpart of the 0.62 agreement, and it is consistent with it.

And the knob does not buy it. This is the part we did not expect. Rank stability is what binds, and it binds hardest exactly where the explanation is largest: 94% of `heart`'s failures at $\lvert S\rvert$=7.9 against 34% of `diabetes_reg`'s at $\lvert S\rvert$=3.1. The mechanism is immediate once stated — a longer prefix is a longer chain of adjacent ranks that all have to hold — and it runs the wrong way for us. Moving from ε_{sel}=ε to 0.35 raises the validity radius on every dataset and *lowers* the identity radius on three of four (0.0107 → 0.0086 on `wine`, 0.0244 → 0.0164 on `diabetes_reg`, 0.0308 → 0.0307 on `abalone`; `heart` is the exception at 0.0017 → 0.0021). So the stability–simplicity frontier trades simplicity for *surrogate validity* and not for explanation stability, and buying features to get the former costs a little of the latter. We would rather state that than leave the title's "stable" to be read as the stronger claim.

Where a tighter version would have to attack. Rank stability. It is far stronger than the result needs, since most permutations do not change the returned pair — but weakening it is not free: within the included set, swapping two features changes which one the greedy adds first and therefore can change the set itself, so a boundary-only condition is unsound. A version that reasoned over all orderings consistent with the intervals, rather than requiring one, is the obvious next step and is not something we have. Artifacts: `results/explanation_stability/identity_certificate_seed42.json` and its seed 43–46 companions (`elasticnam/experiments/run_identity_certificate.py`).

A54 Certified-Stable Explanations: Details

Bound: unsigned (arm A) throughout this appendix, the explainer's own acceptance test; App. A49 gives the arm D counterparts where they exist. Protocol: the small-capacity $(128, 128)$ *encoder with in-sample validation (App. A48).*

Post-hoc stability under certified perturbation, as a table. The body reports these as prose. **This panel is seed** 42 **alone**; the body's figures are five-seed means (0.71, 0.70, 0.56, 0.62), which is why they differ from the column below. On 100 instances across four datasets, mean certified radius 0.23–0.34 — the same seed-42 radii §5.3 quotes for both certificates together; the five-seed means are 0.21–0.31 — every explainer explaining the same VIB-NAM and scored by perturbing that same model's inputs, with the baselines given the same $k = \lvert S\rvert$ our stability-selected explanation produces:

(a) Is the shown explanation still valid?	fraction
Ours — certified	**1.00**
the post-hoc explainers below	*no analogue*[†]
(b) Is the top-k set unchanged?	fraction
exact additive Shapley $f_i(x_i)-\mathbb{E}[f_i]$	0.68
KernelSHAP (200 samples)	0.63
local ridge surrogate (LIME-style (Ribeiro et al., 2016))	0.54
ours, re-selection	0.56

[†]Attribution methods return scores, not a validity predicate, so panel (a) has no counterpart for them. This is *not* a claim that no prior work certifies explanations: formal XAI does, for different objects and model classes (App. A38).

Where the radius goes when you ask for minimality. The feasibility slack remaining at $r=0$ is what a perturbation has to spend, and it is almost entirely consumed by the greedy's stopping rule (five seeds, $G=10^6$, same artifact as the frontier table below):

Table 48: **What minimality spends.** Five seeds, small-capacity in-sample protocol, $G=10^6$, arm A: the slack at $r=0$ is what is left of ε before any perturbation.

Dataset	d	$\lvert S\rvert$	slack at $r=0$	budget used	radius
`abalone`	8	2.5	0.0631	87.4%	0.1332
`wine`	11	3.5	0.0338	93.2%	0.0584
`diabetes_reg`	10	2.8	0.0273	94.5%	0.0670
`heart`	13	8.3	0.0085	98.3%	0.0140

The stability–simplicity frontier. Certifying against $\varepsilon=0.5$ while selecting against ε_{sel}:

Table 49: **The frontier, tabulated.** Five seeds, small-capacity in-sample protocol, arm A, with the censored share beside each mean so that a mean drawn against the ceiling is visible as a lower bound.

Dataset	ε_{sel}	$\lvert S\rvert$	$\lvert S\rvert/d$	mean radius	frac. certifiable	frac. censored
`abalone` (d=8)	0.50	2.5	31%	0.133	0.99	0.03
`abalone`	0.35	3.3	41%	0.383	1.00	0.47
`abalone`	0.20	4.4	55%	0.471	1.00	0.83
`heart` (d=13)	0.50	8.3	64%	0.014	0.93	0.00
`heart`	0.35	9.1	70%	0.324	1.00	0.26
`heart`	0.20	10.1	78%	0.465	1.00	0.76

Five seeds, $G=10^6$. Artifact: `results/explanation_stability/stability_frontier_G1e6_seed42.json` and its four sibling seeds; the superseded $G=10^5$ run is kept as `results/explanation_stability/stability_frontier_seed42.json` and is not cited for any number in this paper.

The simplicity cost is large and we report it as a fraction, not a count. On `heart` the stability-selected explanation already names 64% of the features before any extra stability is bought, and 70% at $\varepsilon_{\text{sel}}=0.35$ (67% on the clean protocol the body quotes). An explanation naming two thirds of the inputs is a weak explanation whatever its certified radius, so the frontier is a real trade and not a free lunch: it is useful where d is small or the model is genuinely sparse, and it degrades toward vacuity as $\lvert S\rvert/d$ approaches 1. The last column is the fraction of instances whose radius hits the precomputed ceiling $r_{\max}=0.5$ and is therefore right-censored: the apparent saturation toward 0.45–0.50 at small ε_{sel} is partly this artifact, not a genuine plateau, and we report it rather than quoting the censored means alone. Artifact: `results/explanation_stability/stability_frontier_G1e6_seed42.json`.

Per-dataset stability under certified perturbation. 25 instances per dataset, 40 perturbations each, drawn uniformly from the certified ball. Every explainer explains the same VIB-NAM and is scored by perturbing that model's inputs; k is set to $|S|$ so all methods select equally many features.

The soundness check is reported separately, on purpose. A column "Ours (valid) $= 1.00$" beside these baselines would not be a comparable measurement: it is the value the certificate *guarantees*, so a table placing it next to measured baselines reads as a comparison when it is a regression test. It is stated as what it is — on all 100 instances the displayed explanation remained valid under every one of the 40 perturbations drawn from its certified ball ($4000/4000$), recomputed from the model and the surrogate coefficients rather than by re-reading the error table. A value below 1.00 would be a soundness bug, not a weaker result. "re-select" below is the strictly stronger property we do *not* claim.

Table 50: Fraction of perturbations on which the top-k feature set is unchanged, with 95% bootstrap CIs. Higher is more stable. Our own certificate is *not* a column here (see text). The instance bootstrap is seeded (`default_rng(12345)`, 10^4 resamples), so these endpoints are reproducible rather than approximately reproducible, and the table gate regenerates them. Artifact: `results/explanation_stability/explanation_stability_seed42.json`.

Dataset	n	radius	k	exact SHAP	KernelSHAP	ridge	re-select
`abalone`	25	0.232	3.1	0.65 [0.53,0.77]	0.67 [0.55,0.77]	0.66 [0.55,0.76]	0.58 [0.48,0.68]
`wine`	25	0.243	3.8	0.67 [0.53,0.80]	0.65 [0.53,0.78]	0.67 [0.56,0.78]	0.64 [0.51,0.78]
`diabetes_reg`	25	0.345	4.4	0.69 [0.60,0.78]	0.65 [0.56,0.74]	0.74 [0.64,0.83]	0.57 [0.46,0.68]
`heart`	25	0.301	8.8	0.70 [0.61,0.79]	0.57 [0.44,0.69]	0.08 [0.04,0.12]	0.46 [0.36,0.56]
pooled, 5 seeds	500	—	—	0.71 [0.66,0.75]	0.70 [0.64,0.76]	0.56 [0.53,0.59]	0.62 [0.56,0.68]

Intervals. The per-dataset rows are seed 42 with 95% instance-bootstrap CIs (10^4 resamples; each instance contributes a rate over its own 40 perturbations). The pooled row replicates the entire protocol on **five training seeds** and reports the between-seed interval, which is the quantity a reader should hold us to: an instance bootstrap answers "what if we had drawn different test points", not "what if we had trained a different model". The KernelSHAP-minus-exact-Shapley gap is 0.8 points at the point estimate and not significantly different from zero under each unit (which does not make it zero): dataset-clustered $[-6.1, 7.6]$, per-dataset $-0.9/-1.6/-1.7/+7.2$ with `heart` carrying it alone, by seed $[-2.1, 3.6]$. **Three things the intervals decide.** (i) exact Shapley $[0.66, 0.75]$ and KernelSHAP $[0.64, 0.76]$ overlap. Overlap is not equivalence, and we set no equivalence margin, so we do not claim the two are equally fragile; the claim the data support is weaker and sufficient: the exact attribution, which samples nothing, is itself fragile, so fragility does not require an estimator. (ii) Our re-selection $[0.56, 0.68]$ overlaps exact Shapley's marginal interval, but the comparison is paired (same models, instances and perturbations) and the paired difference is negative on all five seeds, -0.09 $[-0.12, -0.05]$, so our deficit on set stability is real and not noise. (iii) The ridge surrogate's collapse to 0.08 on `heart` is *seed-specific*: across seeds its pooled figure is 0.56 $[0.53, 0.59]$, far tighter than the $[0.22, 0.73]$ a dataset-cluster bootstrap on seed 42 alone suggested. Replication, not resampling, was what settled that one. Artifacts: `results/explanation_stability/stability_ci_seed42.json`, `results/explanation_stability/stability_seed_variance.json` (`scripts/stability_ci.py`, `scripts/stability_seed_variance.py`).

Reading. The ridge surrogate's collapse on `heart` (0.08) is the sampling variance of a local linear fit in d=13; KernelSHAP is steadier but still flips a third of the time; and the exact additive Shapley value — no sampling at all — flips too (29%), which is the finding that matters: on this model class fragility does not need an estimator; it arises from attributing at a point. Our re-selection column moves at a comparable rate for the same reason, and we do not claim an advantage on this metric. What the certificate changes is that the explanation *already shown* to a stakeholder cannot be invalidated inside the stated radius — a guarantee none of these methods provides. **Soundness note.** $e_i^{\mathcal{C}}$ is a supremum over the error table's tabulated range, so certifying on a ball of radius r requires that range to contain $x_i \pm r$; the table is therefore built with margin $\geq r$ and `explanation_slack` raises if a query would leave it (the table's default anti-rounding margin alone would make the reported radii unsound).

A55 Compression and the Certified Radius (G3): Full Result

Compression enlarges the certified-radius bound (G3). The registered criterion (C5) was: validation-selected β must improve the certified radius by $\geq 15\%$ over $\beta{=}0$ at $\leq 1\%$ relative clean-score loss on ≥ 3 regression datasets, else β is described as a diagnostic only. It is not met, and β is so described. A linearized-Gaussian-channel heuristic — not a theorem; its step from output variance to a Lipschitz constant is unproved — suggests $r_i(\tau) \gtrsim \tau/(|w_i| + L_i^{\mathrm{dec}}\sqrt{1-e^{-2R_i}})$, increasing toward the bypass-only limit $\tau/|w_i|$ as the rate $R_i \downarrow 0$ (Appendix A22); the measurement in App. A57 finds no such slope. Empirically: raising β from 0 to a moderate value lifts the mean certified radius by $+23$–33% on four regression tasks at ≤ 0.025 clean-score cost (Figure 11b). The realized radius is *not* monotone — the optimum is interior ($\beta \in \{0.1, 1, 10\}$) on $8/9$ datasets. **Preregistered outcome (this test fails).** The benefit is *not reliably realizable*: under the validation-selection protocol the 1% clean-score budget selects $\beta{=}0$ on 2–5 of the 9 datasets across the five seeds, and on diabetes the selected $\beta{=}100$ *underperforms* $\beta{=}0$ on test (0.37 vs. 0.45). We therefore present β as a rate–robustness *diagnostic* with an empirically monotone *bound*, not a tuning knob a deployer can bank on. The effect saturates on classification (certified-at-0.1 already ≥ 0.9); the per-dataset, per-β radii behind Figure 11b are the `vibnam_beta*` entries of the certification artifacts (§4). Finally, Eq. (2) gives a certified radius *per feature* — a decomposition black-box verifiers do not provide, a robustness attribution: on heart, per-feature radii span $r_i = 0.05$ to the full $\varepsilon_{\max}{=}1.0$ budget, localizing where a bounded adversary could move the prediction.

A56 Full PGD Robustness Table

Per-dataset PGD-attacked test scores corroborating the exact certification of §4 (the empirical study demoted to corroboration once the exact certificate became the headline).

Table 51: **Robustness within the differentiable class.** $\varepsilon{=}0.20$ PGD-attacked test-set R^2 (regression) or AUROC (classification); 5 seeds, beta selected on the clean metric (no oracle peeking). VIB-NAM wins $7/9$ datasets vs. MLP (paired Wilcoxon over 9 dataset means, $p{=}0.027$ *one-sided* against the preregistered direction; $p{=}0.055$ two-sided. MLP target normalized to match VIB-NAM's $y_{\mathrm{mean}}, y_{\mathrm{std}}$ handling, which costs us one win relative to the $8/9$ of the un-normalised three-split aggregate `results/adversarial/pgd_5seed_summary.json`. The cells below are `results/adversarial/pgd_5seed_all_summary.json`, the nine-dataset family this $7/9$ and p come from, and the table gate regenerates every one of them). EBM is the non-differentiable reference, attacked only by random L_∞-ball perturbation: a weaker attack than PGD, so its column is an optimistic estimate of EBM's worst case and is not comparable to the other two.

Dataset	VIB-NAM (best β)	MLP	Δ (VIB-NAM favor)	EBM (ref.)
heart (cls)	0.736	0.730	$+0.007$	0.893
breast_cancer (cls)	0.936	0.903	$+0.033$	0.988
adult (cls)	0.696	0.566	$+0.130$	0.796
california (reg)	-0.257	-0.243	-0.015	0.245
concrete (reg)	0.089	-0.184	$+0.273$	0.694
bike (reg)	0.310	$+0.087$	$+0.223$	0.697
wine (reg)	0.015	$+0.043$	-0.029	0.273
abalone (reg)	0.055	$+0.004$	$+0.051$	0.478
diabetes_reg (reg)	0.425	$+0.399$	$+0.026$	0.463
Mean	—	—	$\mathbf{+0.078}$	—
Wins	—	—	$\mathbf{7/9}$	—

A57 Negative Finding: IB-Controlled Lipschitz Bound

We initially conjectured a β-controlled certified Lipschitz bound: that for the deterministic predictor $\hat{y}(\mathbf{x}) = \sum_i F_i(x_i)$, the worst-case shift under an L_∞ perturbation $\|\delta\|_\infty \leq \varepsilon$ would satisfy $|\hat{y}(\mathbf{x} + \delta) - \hat{y}(\mathbf{x})| \leq \varepsilon \cdot \sum_i \big(|w_i| + L_g K_i/\sqrt{\beta_i}\big)$, i.e., a Lipschitz constant decaying as $O(1/\sqrt{\beta})$. We tested this prediction by training VIB-NAM with β_i *frozen* at target values $\{0.01, 0.1, 1, 10, 100\}$

(initialising each $\tilde{\beta}_i$ to $\mathrm{softplus}^{-1}(\beta_i)$ and removing $\tilde{\beta}_i$ from the optimiser), then measuring the empirical Lipschitz under PGD ε=0.10 on each of the nine UCI datasets.

P1 (adversarial accuracy monotone in β): partial. Two of nine datasets show a monotone trend ($|\rho_{\log\beta,\,\mathrm{adv}}| > 0.5$, Pearson against the ε=0.20 score): abalone at ρ=+0.88 and diabetes_reg at ρ=−0.51, so only abalone moves in the predicted direction. The other seven peak at an interior β (10 on three, 1 on two, 0.1 on two).

P2 (Lipschitz $\sim 1/\sqrt{\beta}$, log-log slope -0.5): falsified. Mean log-log slope across nine datasets is -0.001; zero datasets within ± 0.1 of the predicted -0.5.

P3 (VIB-NAM Lipschitz $\ll$ EBM Lipschitz): flipped. EBM's empirical Lipschitz is smaller than VIB-NAM's on all nine datasets — by $41\times$ on `heart` and $319\times$ on `breast_cancer`, the two smallest — because its piecewise-constant histogram lookups have zero local Lipschitz inside bins; the margin narrows to $1.4\times$ on `california`. Means over five seeds and the frozen-β grid at PGD ε=0.10; the same population as Table 51.

Mechanistic explanation. At deterministic inference (eval mode $\to$ μ-only path), the prediction $\mathbb{E}_{z\sim q(\cdot|x)}[g(z)]$ marginalises the encoder variance σ. β controls σ alone, so it becomes invisible at the prediction layer. We verified this by re-running PGD with stochastic-encoder MC sampling (N=20); the bound was *still* not tight (mean log-log slope $+0.012$), because β-induced noise lives in latent $\mathbf{z}$ space and the encoder Jacobian intervenes between latent and input perturbations: randomised smoothing in latent space is not randomised smoothing in input space.

The surviving statement. Theorem 1 (graceful degradation) is unaffected by this negative result: it bounds the *shape function form* under high β ($f_i \to w_i x_i + b_i$), not the input-space Lipschitz of the full predictor. The empirical robustness story of §4 (VIB-NAM vs. MLP, $7/9$ wins under target-normalized PGD-20) rests on this graceful-degradation behaviour rather than on the falsified bound.