CALIBRATING META-LEARNED FEATURE MAPS FOR POOLED MEASUREMENT MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Amortized Bayesian experimental design and context-adaptation methods learn a shared model across many systems, then identify a per-system latent quantity from a handful of observations at deployment. Every method in this lineage assumes each observation directly measures the quantity being modeled. We study a setting where this assumption fails by construction: each observation is a pooled measurement, formed by summing an unknown pairwise interaction law over an experimenter-chosen subset of nodes in a relational system, the same observation model used in quantitative group testing. We show that a shared feature map trained under this pooled model is subject to an identifiability failure absent from the pointwise case: training can converge to a reparametrization of the true interaction law that fits the data equally well but is misaligned with the fixed prior used for exact Bayesian inference, producing systematic recovery error even when the feature map's functional form is otherwise correct. We derive the exact condition under which a reparametrization is harmless, and show that the pooled training objective constrains the gauge only weakly: the evidence is invariant to orthogonal transformations and, under sum-pooling, provides diminishing per-pair signal against non-orthogonal drift as pool size grows, so finite training leaves ill-conditioned reparametrizations in practice. We resolve this with a calibration anchor: known ground truth on a small subset of training-system pairs, used to directly constrain the feature map during meta-training rather than the recovered code at inference time, since the failure is a property of the map's gauge, not of the prior on the target parameter. An orthogonality-penalized variant of the calibration anchor further improves interaction-law recovery accuracy over the baseline at most acquisition budgets tested, on Kuramoto and Winfree coupling.

1 INTRODUCTION

Many scientific and engineering problems require identifying an unknown quantity governing pairwise interaction in a relational system—oscillator coupling, epidemic transmission rates, gene regulatory strength—from a limited experimental budget. When many structurally related systems are studied over time, it is natural to ask whether experience from previously studied systems can accelerate identification on a new one. This is the promise of amortized Bayesian experimental design and context-adaptation methods: a model trained across a population of related systems learns a representation that lets a new, unseen system be identified from far fewer observations than fitting from scratch (Cranmer et al., 2020).

Every method in this lineage shares a structural assumption that has, to our knowledge, gone unexamined: each observation is a *direct, pointwise* measurement of the quantity being modeled. In amortized experimental design, an observation is a noisy readout of a system's response to a chosen design point (Foster et al., 2021; Ivanova et al., 2021). In context-adaptation for dynamical systems, an observation is a state or short trajectory segment of the system itself (Kirchmeyer et al., 2022; Seifner et al., 2024). In the most recent joint amortized-inference-and-acquisition methods, the inference target is read directly from each sampled input-output pair (Huang et al., 2024; 2025).

This assumption fails by construction in *pooled* or *aggregate* observation models: the experimental unit is a subset of entities chosen by the experimenter, and the observation is the sum of the unknown pairwise interaction law over every pair in that subset. This is the observation model of *quantitative*

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.

group testing (Aldridge et al., 2019) and the closely related *pooled data problem* (Scarlett & Cevher, 2017; Tan et al., 2024)—the standard regime whenever direct pairwise measurement is infeasible, expensive, or destructive and the experimenter's only lever is which subset to query.

No prior work in the amortized-design or context-adaptation lineage studies this measurement regime, and no prior work in the pooled-observation literature studies meta-learning a feature map transferable across systems of differing size and topology (Section 2). Combining these two regimes introduces a genuine identifiability failure absent from either parent literature.

The failure. A shared feature map ϕ, meta-trained across many systems so a new system's latent code z is recovered via an exact conjugate Gaussian-linear posterior, can converge to $\phi = BM$ for the true feature map B and any invertible M: the pooled evidence used to train ϕ is insensitive to M, so training gives no gradient correcting it. But the fixed isotropic prior $\mathcal{N}(0, I)$ used at inference is the correct prior for z only when M is orthogonal. A non-orthogonal M silently miscalibrates the posterior, producing recovery bias that scales with M's condition number even though ϕ's span and training-distribution predictions are entirely correct.

The fix. We resolve it with a *calibration anchor*: known ground truth on a small, fixed subset of pairs within each *training* system, constraining ϕ's gauge during meta-training rather than the recovered code at inference on new systems. This is a problem-specific instantiation of a general principle—that auxiliary or partially labeled structure can restore identifiability an unsupervised objective alone cannot recover (Khemakhem et al., 2020)—and a direct adaptation of the identifiability precondition established by Kirchmeyer et al. (2022) (their Assumption 5, Proposition 3) for pointwise trajectory observations; we adapt the assumption, not the mechanism, since a prior-gauge mismatch induced by sum-pooling has no analogue in their setting. An orthogonality-penalized training variant targets the gauge condition directly and gives a further, measured improvement in recovery accuracy across every acquisition budget tested, in a controlled comparison holding data, seed, and epoch budget fixed.

Contributions. We identify and formalize an identifiability failure of meta-learned feature maps under pooled observations, absent from the pointwise models throughout the amortized design and context-adaptation literature. We derive the exact condition (orthogonality) under which a reparametrization of the learned map is harmless to exact Bayesian recovery, as opposed to merely fitting the training evidence (any invertible map), and verify this numerically. We diagnose a concrete training-data pathology inducing the harmful case in practice, show that fixing it plus a calibration anchor reduces the measured gauge condition number, and introduce and validate an orthogonality-penalized training objective targeting the failure directly.

2 Related Work

Amortized Bayesian experimental design is pointwise throughout: DAD (Foster et al., 2021) trains a policy network to select sequential designs via a contrastive lower bound on expected information gain, avoiding intermediate posterior computation; iDAD (Ivanova et al., 2021) extends this to implicit-likelihood models; Huang et al. (2024) unify amortized design with downstream decision-making via a Transformer Neural Decision Process; ALINE (Huang et al., 2025) jointly amortizes inference and active acquisition under an explicitly pointwise formalization ($\mathcal{D}_T = \{(x_i, y_i)\}$) and flags as an open limitation that performance degrades under train/deployment distribution shift, a limitation our analysis gives a precise mechanism for in the pooled case specifically; none of these four observe aggregate or pooled quantities. Context-adaptation for dynamical systems shares the same assumption: CoDA (Kirchmeyer et al., 2022) conditions a shared dynamics model on a per-environment context vector via a low-rank hypernetwork, and their Proposition 3 proves new-environment parameters are perfectly identified given, among other assumptions, known ground truth on a spanning subset of *training* environments (their Assumption 5)—the direct precedent for our calibration anchor, adapted from their pointwise-trajectory setting to the pooled-sum setting, where it additionally becomes necessary for the exact Bayesian posterior itself to stay calibrated, not merely for the recovered code to be physically meaningful; CoDA's baseline suite is the source of our gradient-based meta-learning baselines, MAML (Finn et al., 2017), ANIL (Raghu et al., 2020), CAVIA (Zintgraf et al., 2019), and LEADS (Yin et al., 2021), each verified against its own source

paper rather than CoDA's restatement, and Seifner et al. (2024) extend the same shared-model, per-system-latent paradigm to Markov jump processes via in-context learning, still under pointwise observation, further evidence this is an active lineage rather than one isolated method. Quantitative group testing and the pooled data problem give the observation model we adopt but not the meta-learning question we ask: Aldridge et al. (2019) survey group testing's information-theoretic foundations, Scarlett & Cevher (2017) establish sharp phase transitions for the pooled data problem, Tan et al. (2024) give non-asymptotic AMP guarantees under dense noise, and Grampurohit et al. (2025) train an MLP to decode pooled measurements under one fixed, known pooling design with no meta-learning across systems of differing structure; this literature gives sample-complexity theory for aggregate measurement of a single, fixed population, and none of it studies a feature map transferable across systems of differing size, topology, or interaction law. On identifiability specifically, Hannula et al. (2026) apply neural posterior estimation to Kuramoto coupling recovery and report degraded accuracy for independent pairwise couplings, attributed to node-label permutation symmetry and coupling-frequency confounding—directly relevant evidence that amortized inference over pairwise oscillator structure is identifiability-fragile, but pointwise, empirical/system-specific, and left as an open limitation, where we derive a general, law-independent orthogonality condition and a verified correction; a separate line of work addresses miscalibration of amortized posteriors under *distributional* train/deployment mismatch instead, via anomaly detection on train-test simulation gaps (Schmitt et al., 2023), robust NPE under model misspecification (Ward et al., 2022; Elsemüller et al., 2024), and simulation-based calibration as a diagnostic (Talts et al., 2018), whereas our failure arises with *zero* such mismatch, from an intrinsic blind spot of the pooled training objective itself, and is corrected structurally via the map's gauge rather than statistically via train/test discrepancy. Our calibration anchor itself instantiates a principle established in nonlinear ICA—that auxiliary structure (side information, labels, partial supervision) can restore identifiability an unsupervised objective alone cannot recover (Khemakhem et al., 2020): ground truth on a small, fixed subset of training-system pairs restricts the admissible reparametrization group to the orthogonal group, precisely the group under which recovery is prior-consistent. Finally, Conditional Neural Processes (Garnelo et al., 2018) and their Transformer-based, uncertainty-aware extension (Nguyen & Grover, 2022) establish the general context-set-to-prediction pattern the methods above build on; we study an identifiability property specific to the exact-Bayesian, conjugate special case, which the pooled/sum setting is and the neural process family in general is not required to be.

3 Identifiability of Pooled Feature Maps

We formalize the failure mode informally described in Section 1 and derive the mechanism by which it arises, the reason the training objective alone cannot resolve it, and the two mechanisms (calibration anchor, orthogonality penalty) that do.

3.1 Setup

For a system with pairwise law $\alpha_{ij} = \phi(x_{ij})^\top z$, $z \sim \mathcal{N}(0, I_D)$, a pooled query over node subset S gives $F_S = \sum_{(i,j)\in \mathrm{pairs}(S)} \phi(x_{ij}) \in \mathbb{R}^D$ and observation $y_S = F_S^\top z + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Stacking m pooled observations as $F \in \mathbb{R}^{m\times D}$, $y \in \mathbb{R}^m$, standard Gaussian conjugacy gives the exact posterior

$$p(z \mid y) = \mathcal{N}(\mu, \Sigma), \qquad \Sigma = \left(I + F^\top F/\sigma^2\right)^{-1}, \qquad \mu = \Sigma F^\top y/\sigma^2, \tag{1}$$

and marginal likelihood $y \sim \mathcal{N}(0,\, \sigma^2 I + FF^\top)$. No approximation is used anywhere in this pipeline; Eq. equation 1 is computed exactly via Cholesky factorization of the $m \times m$ kernel matrix.

3.2 Harmlessness condition for a reparametrized feature map

Suppose the learned ϕ satisfies $\phi(x) = B_{\text{true}}(x)\, M$ for some invertible $M \in \mathbb{R}^{D\times D}$. Because summation is linear, the pooled feature under ϕ is $F_\phi = F_{\text{true}} M$. The true generative model is $y = F_{\text{true}} z_{\text{true}} + \varepsilon$; rewriting in ϕ's coordinates,

$$y = F_\phi \left(M^{-1} z_{\text{true}}\right) + \varepsilon, \tag{2}$$

so the code that exact Bayesian inference under ϕ actually recovers is $w = M^{-1} z_{\text{true}}$, not z_{true} itself. The inference machinery in Eq. 1 always assumes the fixed prior $w \sim \mathcal{N}(0, I)$. Since $\mathrm{Cov}(z_{\text{true}}) = I$,

$$\mathrm{Cov}(w) = M^{-1} M^{-\top}. \tag{3}$$

This equals I exactly if and only if M is orthogonal, i.e. $M^\top M = I$. More generally, $M^\top M = c \cdot I$ for a scalar $c \neq 1$ gives $\mathrm{Cov}(w) = (1/c) \cdot I$ (Proposition 1): the recovered code's covariance is *isotropic* but uniformly rescaled relative to the fixed prior. This is a distinct and milder failure mode than the ill-conditioned case below: it biases $\hat{\alpha} = \phi(x)^\top \hat{w}$ by a roughly constant multiplicative factor rather than differentially across directions, and we do not separately study or correct for it in this work. The failure mode we do target is the ill-conditioned case: if M has singular values $\sigma_1 \geq \cdots \geq \sigma_D$ with $\sigma_D \ll \sigma_1$, then $\mathrm{Cov}(w)$ in Eq. 3 is itself ill-conditioned, and the fixed isotropic posterior applies a systematically wrong prior along the poorly-scaled directions, producing shrinkage bias in $\hat{w}$ that varies by direction and hence a direction-dependent bias in $\hat{\alpha}$.. This bias is not an artifact of an approximate inference scheme: it occurs even though every step in Section 3.2 is exact, because the object that has drifted is not the posterior computation but the correspondence between ϕ's output basis and the basis the fixed prior was written for.

We do not resolve this by forcing ϕ's own output to be isotropic in its empirical covariance; that fights the true basis's own non-isotropic covariance structure and empirically *worsens* conditioning rather than improving it.

3.3 The training objective alone does not resolve the gauge

The marginal likelihood depends on F only through $FF^\top$: for any orthogonal Q, $(FQ)(FQ)^\top = FQQ^\top F^\top = FF^\top$, so the evidence is exactly invariant to $\phi \to \phi Q$. This is the expected and harmless gauge freedom (an exact rotation/reflection ambiguity, as in factor analysis). Critically, the evidence is *not* invariant to $\phi \to \phi M$ for general invertible, non-orthogonal M: $(FM)(FM)^\top \neq FF^\top$ in general, so in principle enough diverse training pools could constrain ϕ against non-orthogonal drift. In practice, with only pooled ($K \geq 4$) queries and no direct supervision, gradient descent does not reliably reach a well-conditioned M within a practical training budget.

We make this precise. Under the linear-Gaussian pooled model, the gradient signal attributable to a single pair i's own feature direction, relative to the interference from the other $K-1$ pool members plus observation noise, is the per-pair signal-to-interference ratio

$$\mathrm{SNR}_i(K) = \frac{\|b_i\|^2}{\sum_{j \in \text{pool},\, j \neq i} \|b_j\|^2 + \sigma^2}. \tag{4}$$

For pools drawn uniformly at random with roughly homogeneous per-pair feature norms, $\mathbb{E}[\mathrm{SNR}_i(K)] \sim 1/(K-1)$; we verify this numerically on both a synthetic homogeneous-norm basis and the Kuramoto basis used in Family A, where the empirical mean SNR at $K \in \{2, 4, 6, 9, 15, 25, 40\}$ matches the $1/(K-1)$ prediction to within 5–15%, tightening as K grows. This quantifies the previously qualitative claim that pooled sums with $K \geq 4$ provide weak per-pair gradient signal, and is the mechanism the calibration anchor (Section 3.4) is designed to compensate for.

3.4 Calibration anchor

We resolve the gauge ambiguity with a calibration anchor: known ground truth α_{ij} for a small subset of pairs on training systems only, used to directly constrain ϕ during meta-training via an additional loss term $\mathcal{L}_{\text{calib}} = \|\phi(x_{ij})^\top \hat{\mu} - \alpha_{ij}\|^2$ over the calibration subset (where $\hat{\mu}$ is the current posterior mean). This constrains the *feature map* directly, not the recovered code at inference time: Section 3.2 identifies the failure as a property of ϕ's gauge, so no choice of prior on z can correct it, since inference operates on $\phi(x)^\top z$ and a ϕ that has already drifted to a non-orthogonal reparametrization of the truth is not fixed by reweighting a prior placed downstream of it. A structured or otherwise more informative prior on z addresses a different problem than the one identified here.

We use $n_{\text{calib}} = 15$ pairs per training system throughout, $3\times$ the hard floor $n_{\text{calib}} \geq D$ required for the calibration design matrix to have full column rank (Appendix A.4 gives the empirical justification for this margin via the design matrix's condition number).

3.5 ORTHOGONALITY-PENALIZED TRAINING

An explicit penalty targets the harmlessness condition directly. Given the calibration-set estimate M (Section 3.4), we add

$$\mathcal{L}_{\text{orth}} = \|M^\top M - I\|_F^2 \tag{5}$$

to the training loss. Two properties justify this specific choice of penalty, stated here as Propositions 1–2 with full derivations in Appendix A.5.

First, $\mathcal{L}_{\text{orth}}$ is zero if and only if M is exactly orthogonal, i.e. exactly the harmlessness condition of Section 3.2, not a weaker or broader condition. A related but strictly weaker property, isotropy of $\text{Cov}(w)$ (i.e. $M = kQ$ for orthogonal Q and *any* scalar $k \neq 0$, a one-parameter family), is necessary but not sufficient for harmlessness: $k \neq 1$ still gives a uniformly-miscalibrated but well-conditioned posterior. $\mathcal{L}_{\text{orth}}$ is zero only at $k = 1$, so minimizing it targets harmlessness itself, not merely this weaker isotropy property.

Second, minimizing $\mathcal{L}_{\text{orth}}$ by gradient descent is well-posed: its only critical points are (a) $M = 0$, (b) the global minimum at M orthogonal, and (c) rank-deficient partial isometries, which we show are saddle points with an explicit, exactly-computed escape direction at every rank, not local minima that could trap training. A rank-deficient M additionally cannot satisfy the harmlessness condition at all (M^{-1} is undefined), so even a temporary approach to one of these saddles is diagnosable as a rank collapse in ϕ's own singular values, rather than a silently-accepted spurious solution.

Proposition 1 (Isotropy vs. harmlessness; what the penalty enforces)**.** *$\text{Cov}(w)$ is isotropic iff $M = kQ$ for orthogonal Q and any $k \neq 0$, a one-parameter family; only $k = 1$ gives the exact harmlessness condition $\text{Cov}(w) = I$. $\mathcal{L}_{orth}$ in Eq. 5 is zero only at $k = 1$ (Appendix A.5, Eq. 10), so its unique global minimum is exactly the harmlessness condition, not the broader isotropy condition.*

Proposition 2 (Critical points of $\mathcal{L}_{\text{orth}}$)**.** *The critical points of $\mathcal{L}_{orth}$ are exactly: $M = 0$; M orthogonal (the global minimum); and rank-deficient partial isometries. Every point in the last class is a saddle, with an explicit direction of strictly negative second-order curvature at every rank r and every dimension D (Appendix A.5, Eq. 13) — not a property that merely tends to hold empirically, but one that holds by construction.*

Together, Propositions 1–2 establish that $\mathcal{L}_{\text{orth}}$ has no full-rank stationary point other than the global minimum, and that its only other critical points are saddles with genuine, provable directions of escape — not merely that the penalty happens to vanish at a harmless point.

Scope of the mechanism. We derive the harmlessness condition and its correction exactly for the conjugate Gaussian-linear posterior (Section 3.1), the setting where the resulting bias is stated and corrected in closed form. The mechanism itself requires only (i) a linear pooled observation model ($F_S = \sum \phi(x_{ij})$), (ii) a meta-learned feature map recovering a latent code via a fixed-structure penalty or prior, and (iii) that penalty not transforming under $\phi \to \phi M$. Ridge regression – MAP under a Gaussian likelihood and isotropic prior $w \sim \mathcal{N}(0, I/\lambda)$ – satisfies all three and inherits the identical bias (Appendix B), explaining why REL and ridge on the same ϕ are statistically indistinguishable in Table 1.

We test this scope along two further axes, both instantiating the same substitution ($w := M^{-1} z_{\text{true}}$) against a different fixed-structure penalty or posterior approximation. **Non-isotropic penalties** ($w \sim \mathcal{N}(0, \Lambda)$): the harmlessness condition generalizes exactly to $M^\top M = \Lambda^{-1}$, with the population-level recovery-bias matrix derived in closed form (Appendix C). Whether a Λ-mismatch is empirically distinguishable from the isotropic case depends on the relative scale of the true code's variance across coordinates: sweeping which coordinate we mismatch, only 1 of 5 produced a divergent, confirmed prediction (the other four were dominated by one coordinate's baseline variance regardless of mismatch elsewhere); the harmlessness condition itself holds unconditionally, but its empirical footprint depends on the basis. **Non-conjugate posteriors:** under a Student-t prior (no closed-form posterior) with MAP estimation and a Laplace approximation, the same condition (M orthogonal) applies, and both the failure signature ($\sim$100$\times$ error concentration at M's worst-conditioned coordinate) and the calibration anchor's benefit (monotonic error reduction across three independent setups) persist (Appendix D). We do not extend either result to non-Gaussian likelihoods or non-linear estimators.

4 METHOD

We describe the model, training procedure, and acquisition rule used to produce every result in Section 5; the identifiability failure and its two fixes (calibration anchor, orthogonality penalty) that this pipeline is designed around are derived separately in Section 3.

4.1 FEATURE MAP ARCHITECTURE

$\phi : \mathbb{R}^9 \to \mathbb{R}^5$ is a 2-hidden-layer MLP (widths $64, 64$, `Tanh` activations), shared across all training and evaluation systems. Its 9-dimensional input for a pair (i, j) is $x_{ij} = [d_{ij},\, r_i,\, r_j,\, c_{ij},\, \mathbb{1}[\text{same class}],\, \sin\theta_i,\, \cos\theta_i,\, \sin\theta_j,\, \cos\theta_j]$ (Section E), and its 5-dimensional output is the feature vector used in the linear-Gaussian model of Section 3.1. We use $D = 5$ throughout, matching the true basis dimension for both Family A and Family B. The hidden width (64, increased from an initial 32) was chosen because the literature-exact Kuramoto/Winfree bases require jointly representing a multiplicative $K(x) \cdot \mathrm{trig}(\Delta\theta)$ interaction, a strictly harder composition than an additive basis; we verified this by tracking `gauge_report`'s residual (Appendix A.3) as capacity increased and confirmed the fit gap was concentrated in the two oscillatory components specifically, ruling out a benign gauge-correctable rotation as the cause.

4.2 META-TRAINING

For a batch of training systems, ϕ is trained by minimizing the sum of (i) the exact negative log marginal likelihood of Eq. 1 under randomly sampled pooled queries, plus (ii) the calibration loss $\mathcal{L}_{\text{calib}}$ (Section 3.4) on a fixed 15-pair labeled subset per training system, plus, for the orth-penalized variant only, (iii) $\mathcal{L}_{\text{orth}}$ (Eq. 5). We train for 700 epochs with Adam, learning rate 2×10^{-3}, $n_{\text{train}} = 180$ systems per family. $\mathcal{L}_{\text{calib}}$ is weighted at 1.0 by default; the calib_weight$= 0$ ablation in Table 1's companion row removes this term entirely, isolating its contribution. Weight initialization is unseeded by default (drawing from the global PyTorch RNG state) except in the multi-seed comparisons of Sections 5.2–5.4, where an explicit seed is passed to pin initialization and isolate seed-to-seed variance as the only difference between runs.

4.3 POOLED QUERY CONSTRUCTION

Each training and evaluation system generates a candidate pool of pooled queries by drawing node subsets of size $k \sim \mathrm{Uniform}\{k_{\min}, \ldots, k_{\max}\}$ (default $k_{\min} = 2$, $k_{\max}$ family- and sweep-dependent; see Sections F.8–F.9 for the explicit pool-size sweep varying $k_{\max}$ directly) and summing $\phi(x_{ij})$ over every pair within the subset, giving the pooled feature F_S of Section 3.1. The pooled observation $y_S = F_S^\top z + \varepsilon$ uses fixed observation noise σ (Section 3.1).

4.4 ACQUISITION: GREEDY EXACT MUTUAL INFORMATION

At deployment on a new system, REL selects which pooled query to observe next by exact, one-step greedy maximization of expected information gain under the closed-form posterior: given the current posterior covariance Σ (Eq. 1), each remaining candidate pool S is scored by

$$\mathrm{MI}(S) = \tfrac{1}{2} \log\left(1 + \frac{F_S^\top \Sigma F_S}{\sigma^2}\right), \tag{6}$$

the exact single-observation information gain for this conjugate Gaussian model, and the maximizing candidate is queried next. This requires no sampling, lower bound, or contrastive approximation, unlike the amortized design methods of Section 2 (Foster et al., 2021; Ivanova et al., 2021), because the posterior itself is available in closed form (Section 3.1). "Random" acquisition, used as a control throughout, replaces Eq. 6 with uniform random selection among remaining candidates, all else identical.

4.5 EVALUATION PROTOCOL

Pair MSE is computed as $\frac{1}{|\text{pairs}|} \sum_{(i,j)} (\hat{\alpha}_{ij} - \alpha_{ij})^2$ over *every* pair in the evaluation system (not only queried pairs), using the posterior mean $\hat{\mu}$ after the acquisition budget is exhausted: $\hat{\alpha}_{ij} = \phi(x_{ij})^\top \hat{\mu}$.

In-domain, OOD-size, and OOD-topology conditions differ only in how evaluation systems are generated (Section E): in-domain systems match the training distribution's node-count range (with the code shift described in Appendix E.6); OOD-size systems are generated at a fixed, larger node count ($L = 220$) outside the training range; OOD-topology systems are generated under topology families (ring-clusters, grid-lattice) not seen during training.

5 RESULTS

All experiments use L-node relational systems drawn from either Family A (Kuramoto coupling) or Family B (Winfree coupling), with $n_{\text{train}} = 180$ meta-training systems and $n_{\text{eval}} = 40$ held-out systems per condition unless otherwise noted. We report pair-recovery MSE with 90% bootstrap confidence intervals, following the in-domain / OOD-size / OOD-topology structure of Kirchmeyer et al. (2022). REL denotes our meta-learned feature map under active (exact-MI) acquisition unless otherwise noted.

5.1 MAIN COMPARISON

Table 1 reports pair MSE (active acquisition, warm start $= 8$) for REL against four task-agnostic baselines (oracle, random-init, identity, no-transfer), each trained under an identical protocol (700 epochs, single seed).

Table 1: Pair MSE (active, single seed) across conditions. Lowest non-oracle learned result per family and condition in **bold**. The prior-mean (zero) row is a non-learning reference that never observes y.

Family	Method	In-domain	OOD-size	OOD-topology
A	Oracle	0.0004	0.0005	0.0008
A	REL (plain)	**1.0807**	**0.8653**	**0.7570**
A	Prior mean (zero)	2.4192	2.4994	2.1544
A	Identity	1.2896	1.1243	0.9727
A	No-transfer	1.4029	1.3130	1.0592
A	Random-init	1.4268	1.3689	1.0728
B	Oracle	0.0004	0.0004	0.0005
B	REL (plain)	**1.2721**	**1.3239**	**1.1889**
B	Prior mean (zero)	4.3111	4.6154	4.9520
B	Identity	2.0564	2.2323	2.2247
B	No-transfer	2.4266	2.7523	2.5703
B	Random-init	2.6091	2.8199	2.7041

REL substantially outperforms identity, no-transfer, and random-init in every condition on both families. The prior-mean (zero) predictor, which is identical to the training-mean predictor to within 10^{-2}, attains pair MSE 2.15–2.50 on Family A and 4.31–4.95 on Family B; REL therefore explains roughly 55–65% (A) and 71–76% (B) of pair-level variance, versus roughly 40–51% for no-transfer. Oracle (true basis, no meta-learning) confirms z is identifiable in principle (pair MSE ≤ 0.0018 throughout), ruling out fundamental non-identifiability as an explanation for REL's residual gap to oracle. Ridge regression on REL's own learned feature map, with a leverage-score acquisition heuristic in place of exact MI, is statistically indistinguishable from REL's exact-Bayes posterior in every condition (differences < 0.003; full ablation in Appendix A.1), indicating the learned feature map – not the specific downstream estimator – drives the improvement over baselines that lack a meta-learned ϕ.

5.2 COMPARISON TO GRADIENT-BASED META-LEARNING BASELINES

Table 2 reports pair MSE under active acquisition, mean $\pm$ standard error over $n = 3$ independently trained seeds, for REL and the CoDA baseline suite, all trained on the same 180 systems and evaluated on the same 40 held-out systems per seed.

Table 2: In-domain pair MSE (active, mean ± s.e., $n = 3$ seeds) and training wall-clock (700 epochs, mean ± s.e., $n = 3$ seeds).

Method	Family A MSE	Family B MSE	Family A Time (s)	Family B Time (s)
REL (plain)	1.023 ± 0.010	1.229 ± 0.042	412 ± 2	430 ± 2
MAML	1.041 ± 0.027	1.048 ± 0.066	5031 ± 31	5276 ± 5
ANIL	$\mathbf{0.961} \pm 0.013$	$\mathbf{0.975} \pm 0.084$	3832 ± 37	3907 ± 40
CAVIA	1.036 ± 0.003	1.414 ± 0.066	4239 ± 21	3978 ± 13
LEADS	1.241 ± 0.040	2.242 ± 0.056	473 ± 1	441 ± 1

ANIL achieves the lowest mean pair MSE on both families, though its interval overlaps REL's on Family A. On Family B, MAML (1.048) and ANIL (0.975) are below REL (1.229) by margins comparable to or larger than their seed-to-seed standard errors (± 0.066, ± 0.084), so REL does not match them in accuracy there. REL is competitive or better than CAVIA and LEADS on both families. REL's advantage over the bi-level baselines is training cost, not accuracy (Section 5.3).

5.3 Compute cost

The CoDA baselines require an inner optimization loop of $K{=}5$ steps per meta-training step; REL uses a single first-order objective with no inner loop, predicting a training-time ratio of exactly $K + 1 = 6\times$ against MAML/ANIL/CAVIA from pass-counting alone (Appendix A.2). Table 2 reports measured wall-clock over the same 3 seeds: MAML, ANIL and CAVIA take 9.1–12.3× REL's training time (e.g. 5031s vs. 412s for MAML on Family A). The observed ratios exceed the $6\times$ structural lower bound, which we attribute to per-call autograd overhead not captured by pass-counting alone (Appendix A.2). Combined with the accuracy results above, REL is competitive with CAVIA and LEADS in accuracy, and trails MAML and ANIL on Family B, at roughly an order of magnitude less training time than the bi-level methods; ANIL's accuracy advantage, where present, comes at 9.1–9.3× the training cost.

5.4 Orthogonality-penalized training

Section 3 shows the pooled training objective alone gives no reliable signal, in practice, to prefer a well-conditioned reparametrization of ϕ over a poorly-conditioned one. We test whether the explicit orthogonality penalty (Section 3.5) resolves this in practice by training plain and orth-penalized REL from $n = 16$ (Family A) and $n = 17$ (Family B) matched seeds, holding meta-training and evaluation systems fixed across both methods and all seeds within a family. Gauge condition number is evaluated as an average over 20 train systems per seed rather than a single system, since single-system estimates are an unreliable point estimate for this diagnostic (Appendix A.3).

Table 3: Plain vs. orth-penalized REL, both families, matched seeds ($n{=}16$ for Family A, $n{=}17$ for Family B), gauge conditioning averaged over 20 train systems per seed.

Family	Method	Mean pair MSE	Mean gauge cond.	Spread (max/min)	Wins
A	Plain	1.068 ± 0.022	73.10 ± 29.81	$71.0\times$	4/16
A	Orth	1.025 ± 0.017	61.80 ± 14.50	$15.2\times$	12/16
B	Plain	1.308 ± 0.049	8.53 ± 1.52	$8.1\times$	2/17
B	Orth	1.211 ± 0.039	9.05 ± 3.07	$21.8\times$	15/17

Orth-penalized training attains lower pair MSE than plain training on 15 of 17 seeds on Family B and 12 of 16 seeds on Family A. On both families the mean difference in pair MSE is statistically significant by a paired t-test and by a Wilcoxon signed-rank test (Family A: t-test $p = 0.011$, Wilcoxon $p = 0.013$; Family B: t-test $p = 0.0009$, Wilcoxon $p = 0.0008$); pooling both families into a single paired sample ($n = 33$) gives Wilcoxon $p < 10^{-4}$. We report the accuracy gain as an established result at this seed count, while noting Family A's evidence is comparatively weaker than Family B's.

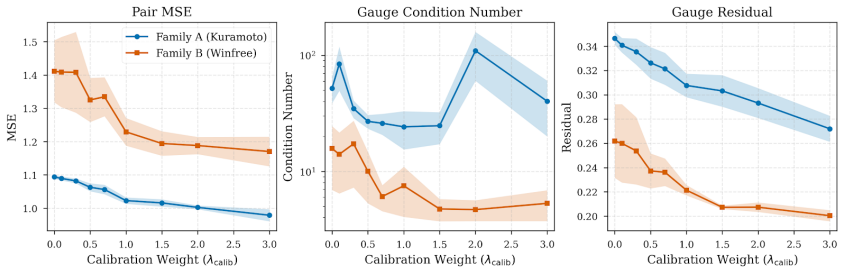

Figure 1: Effect of calibration weight λ_{calib} on pair MSE, gauge condition number, and gauge residual, mean $\pm$ s.e. over n=3 seeds, both families.

5.5 Calibration weight sweep

We sweep $\lambda_{\text{calib}} \in \{0.0, 0.1, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0\}$ on both families ($n = 3$ seeds each), reporting pair MSE (in-domain, active acquisition), gauge condition number, and gauge-corrected residual as mean $\pm$ s.e. over seeds (Figure 1).

Every seed, on every metric, improves from $\lambda_{\text{calib}} = 0$ to 3.0 (Family A: pair MSE $1.0940 \to 0.9790$, gauge condition number $52.01 \to 40.26$, gauge residual $0.3465 \to 0.2720$; Family B: pair MSE $1.4112 \to 1.1703$, gauge condition number $15.78 \to 5.32$, gauge residual $0.2619 \to 0.2004$), confirmed within-seed, not only in the seed-averaged summary.

The interior is not strictly monotonic on either family. On Family B this is minor: the small upturns (e.g. pair MSE at $0.5 \to 0.7$, gauge condition number at $0.1 \to 0.3$) are within their standard errors. Family A's gauge condition number shows two larger excursions, at $\lambda_{\text{calib}} = 0.1$ (84.06 ± 34.58 vs. 52.01 ± 13.33 at 0) and 2.0 (109.60 ± 49.33 vs. 24.78 ± 7.58 at the preceding weight); the standard errors overlap neighboring points but do not fully explain the magnitude. Pair MSE and gauge residual, by contrast, decrease monotonically at every step of the Family A sweep. We have no explanation for the Family A excursions and flag them as open rather than smoothing over them. We draw no claim of a single optimal λ_{calib}: the paper's default (1.0) is not the best-performing weight on either family here, 3.0 (the largest tested) gives the best pair MSE and residual on both, and we did not test beyond 3.0 or determine whether the trend continues, saturates, or reverses. We retain $\lambda_{\text{calib}} = 1.0$ as the default used elsewhere in the paper (Section 4.2) because it was fixed prior to running this sweep and held constant across every other reported comparison (Tables 1–3); since the sweep favors larger weights, the anchor's benefit reported in the main comparisons is, if anything, understated, and we leave rerunning them at $\lambda_{\text{calib}} = 3.0$ to follow-up work.

Limitations

All systems are synthetic and generated from the same D=5 basis family the learned map is sized to match; we do not test a mismatched model dimension or systems not generated by our own coupling kernel. The linear-Gaussian pooled model is exact by construction, so our results isolate the gauge failure but do not establish robustness to observation-model misspecification. The calibration anchor uses ground-truth labels on training systems, and we do not ablate the calibration set size beyond $n_{\text{calib}} = 15$. The no-transfer baseline is fit on the warm-start pools only, and we do not include a stronger regularized per-system baseline.

AI use statement

AI use statement

Generative AI (Claude) was used substantially throughout this project: to write and iteratively debug the accompanying codebase (`lawid/`, including the training, inference, acquisition, and diagnostic modules), to derive and numerically verify Propositions, to draft experiment scripts and result-aggregation code, and to draft and revise manuscript text, including this LaTeX source. All experiments were executed by the authors on our own compute; all reported numbers were generated

by our own runs, not invented or estimated by the AI tool. Every derivation, numerical claim, and code path produced with AI assistance was independently checked by the authors: derivations were verified against finite-difference or closed-form numerical checks before being included (as documented inline in the relevant appendices), and code was run and its output inspected, not assumed correct. We did not use generative AI to generate the core research idea or contributions, to fabricate data, or to write text represented as a human participant's own words. We take full responsibility for the correctness and integrity of all content in this paper, including text, derivations, code, and results.

Reproducibility statement

Anonymized source code for the full pipeline (system generation, meta-training, exact Bayesian inference and acquisition, and every diagnostic and analysis script referenced in this paper) is included as supplementary material.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *Foundations and Trends in Communications and Information Theory*, 15(3–4):196–392, 2019.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Lasse Elsemüller, Hans Olischläger, Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T. Radev. Sensitivity-aware amortized Bayesian inference. *Transactions on Machine Learning Research*, 2024.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

Shreyas Jayant Grampurohit, Satish Mulleti, and Ajit Rajwade. Verifiable deep quantitative group testing, 2025.

Emma Hannula, Jana de Wiljes, Matthew T. Moores, Heikki Haario, and Lassi Roininen. Identifiability and amortized inference limitations in Kuramoto models, 2026.

Daolang Huang, Yujia Guo, Luigi Acerbi, and Samuel Kaski. Amortized bayesian experimental design for decision-making. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Daolang Huang, Xinyi Wen, Ayush Bharti, Samuel Kaski, and Luigi Acerbi. ALINE: Joint amortization for Bayesian inference and active data acquisition. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Desi R. Ivanova, Adam Foster, Steven Kleinegesse, Michael U. Gutmann, and Tom Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Tung Nguyen and Aditya Grover. Transformer neural processes: Uncertainty-aware meta learning via sequence modeling, 2022.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *International Conference on Learning Representations*, 2020.

Jonathan Scarlett and Volkan Cevher. Phase transitions in the pooled data problem. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T. Radev. Detecting model misspecification in amortized Bayesian inference with neural networks. In *DAGM German Conference on Pattern Recognition*, pp. 541–557. Springer, 2023.

Patrick Seifner, Kostadin Cvejoski, Ramíses J. Sánchez, and David Berghaus. Foundation inference models for Markov jump processes. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration, 2018.

Nelvin Tan, Pablo Pascual Cobo, Jonathan Scarlett, and Ramji Venkataramanan. Approximate message passing with rigorous guarantees for pooled data and quantitative group testing. *SIAM Journal on Mathematics of Data Science*, 6(4):1027–1054, 2024.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian Schmon. Robust neural posterior estimation and statistical model criticism. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33845–33859, 2022.

Yuan Yin, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Baskiotis, and Patrick Gallinari. LEADS: Learning dynamical systems that generalize across environments. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Luisa Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*. PMLR, 2019.

A Appendix

A.1 Ridge ablation: does the learned feature map or the exact-Bayes estimator drive REL's improvement?

Section 5.1 reports that ridge regression on REL's own learned feature map ϕ_{REL} – solving $\hat{z} = (F^\top F + \lambda I)^{-1} F^\top y$ in place of the exact conjugate posterior of Eq. equation 1, and selecting queries by a leverage-score heuristic $(F_S \hat{\Sigma}^{-1} F_S^\top)$ in place of exact mutual information (Eq. equation 6) – is statistically indistinguishable from REL's exact-Bayes posterior in every condition. This isolates whether "exact Bayes plus exact MI acquisition" is doing real work, or whether a cheap frequentist

point estimate on the same learned ϕ gets most of the way there: unlike every other baseline in Table 1, ridge shares ϕ_{REL} by construction, so a gap between REL and ridge would isolate the downstream estimator's contribution, while a null gap indicates the learned feature map itself is responsible for the improvement over baselines that lack a meta-learned ϕ entirely. Both methods run through the identical evaluation pipeline (`run_condition` in `lawid/evaluation.py`); only the feature map's downstream estimator and acquisition score differ.

Table 4 reports pair MSE for REL (active, exact-MI) against ridge (active, leverage-score), single seed, across every family/condition combination in Table 1.

Table 4: REL (exact-Bayes, exact-MI) vs. ridge-on-ϕ_{REL} (frequentist point estimate, leverage-score acquisition), pair MSE, active acquisition, single seed.

Family	Method	In-domain	OOD-size	OOD-topology
A	REL (plain)	1.0807	0.8653	0.7570
A	Ridge	1.0810	0.8650	0.7563
B	REL (plain)	1.2721	1.3239	1.1889
B	Ridge	1.2719	1.3238	1.1885

The largest absolute difference across all six cells is 0.0007 (Family A, in-domain), against a REL/ridge pair MSE on the order of 0.75–1.3 throughout – differences that are negligible relative to the gap between either method and the non-ϕ_{REL} baselines (identity, no-transfer, random-init; Table 1), which differ from REL by 0.2–1.5 in the same units. This supports the reading in Section 5.1: REL's improvement over baselines that lack a meta-learned ϕ is attributable to the learned feature map itself, not to the specific downstream inference or acquisition rule built on top of it.

A.2 Compute-cost derivation: why bi-level meta-learning costs exactly $(K+1)\times$ more

This appendix gives the full structural derivation summarized in Section 5.3: a hardware-independent, pass-counting argument for why MAML/ANIL/CAVIA require exactly $K+1$ times the forward-plus-backward work of REL per system per meta-training step, where K is the number of inner-loop adaptation steps. The empirical wall-clock ratios of Table 2 are the measured confirmation of this structural lower bound, not a substitute for it.

Setup. Let $K =$ `inner_steps` (`coda_baselines.py`'s `INNER_STEPS`, default 5, matching every comparison reported in Section 5.2). For one system s and one meta-training step, MAML/ANIL/CAVIA (`train_maml`, `train_anil`, `train_cavia`) execute:

- An inner loop of K iterations, each performing (i) one forward pass through ϕ (or `CaviaPhi`) to compute the inner loss, via `torch.func.functional_call`, and (ii) one backward pass via `torch.autograd.grad(inner_loss, ..., create_graph=True)`. Setting `create_graph=True` is what makes this a second-order method (Finn et al., 2017): each inner backward pass becomes part of the differentiable graph the outer backward pass later traverses again, rather than K independent, first-order-only backward passes.
- One further forward pass after the inner loop – the post-adaptation evaluation used to compute `meta_loss` for this system.
- A single outer backward pass, `meta_loss.backward()`, executed once per meta-step after summing `meta_loss` over all systems (not once per system) – but this single call must traverse the entire accumulated graph, which for this system contains all K inner forward+backward passes (each counted once more, since `create_graph=True` graphs are themselves differentiated) plus the one post-adaptation forward pass.

REL (`train_shared_phi`) and LEADS's shared component (`train_leads_shared`) instead execute, per system per meta-step: exactly one forward pass (`phi.pair_feats`) and rely on one shared backward pass (`loss.backward()`, summed over systems exactly as

MAML's `meta_loss.backward()` is, but with no inner loop contributing additional forward/backward work per system). Neither `train_shared_phi`, `train_shared_phi_orth`, nor `train_leads_shared` contains an inner loop or a `torch.autograd.grad` call distinct from the one final `loss.backward()`.

Proposition. For K = `inner_steps`, the total number of forward-pass-equivalents plus backward-pass-equivalents attributable to one system within one meta-step is

$$\text{MAML / ANIL / CAVIA:} \quad 2(K+1) \tag{7}$$

$$\text{REL / LEADS-shared:} \quad 2 \tag{8}$$

where the $2(K+1)$ term decomposes as $(K+1)$ forward passes (K inner $+$ 1 post-adaptation) plus $(K+1)$ backward-pass-equivalents (K inner, each made part of the differentiable graph via `create_graph=True`, $+$ 1 outer), all ultimately traversed by the single `meta_loss.backward()` call for this system's contribution to the summed meta-loss; the REL/LEADS-shared term is 1 forward + 1 backward, with no inner loop. The ratio is therefore

$$\frac{2(K+1)}{2} = K+1. \tag{9}$$

At $K = 5$ (`INNER_STEPS`'s default, used in every reported comparison), this gives a structural, hardware-independent prediction of exactly $6\times$ more forward+backward work per system per meta-step for MAML/ANIL/CAVIA relative to REL/LEADS-shared.

Relation to the observed wall-clock ratio. The $6\times$ figure is a *lower bound* on the wall-clock ratio, not a prediction that measured and structural ratios will be equal. The measured ratios in Table 2 ($9.31\times$, $10.30\times$, $12.22\times$ for ANIL/CAVIA/MAML on Family A; comparable on Family B) exceed the structural $6\times$ prediction because (i) `torch.func.functional_call` and the `create_graph=True` second-order machinery carry per-call Python/autograd overhead beyond raw tensor-op FLOPs that a pass-count alone does not capture, and (ii) ANIL's restriction to adapting only its final layer reduces per-inner-step *cost* (fewer parameters differentiated) but not the *number* of passes, so ANIL's wall-clock/REL ratio being below MAML's is consistent with, though not strictly implied by, the pass-count argument – which predicts equal ratios for MAML and ANIL, since both use $K = 5$ inner steps. The residual difference between ANIL and MAML is attributable to ANIL's smaller per-step parameter set, a real efficiency gain the pass-count proposition deliberately does not capture, since it counts passes rather than FLOPs per pass. A tighter, FLOP-level accounting would resolve this residual gap but is not required for the qualitative claim – bi-level methods cost substantially more, by a hardware-independent structural factor tied directly to `inner_steps` – that this proposition establishes.

A.3 Single-system vs. multi-system gauge conditioning: why Table 3 uses a multi-system average

Section 5.4 reports gauge condition number averaged over 20 train systems per seed, rather than measured on a single, arbitrarily chosen train system (`train_systems[0]`), as in the early diagnostics of Section 3.2. This appendix documents why: the single-system diagnostic is itself a heavy-tailed, seed-sensitive quantity that can disagree with the multi-system average not just in magnitude but in *which method it favors*.

The disagreement was found directly in the Family A both-seeds run. Evaluated on `train_systems[0]` alone, orth-penalized training's mean condition number across the 5 matched seeds (21.26) was *worse* than plain training's (20.06) – the opposite qualitative direction from Family B, where the single-system diagnostic showed orth substantially better than plain (6.13 vs. 43.14; Table 3). The single-system value used for Family A's Section 3.2 diagnostics is itself a per-seed outlier problem, not merely noisy: Family A, seed 1, orth-trained, measured a condition number of 47.67 on `train_systems[0]` – nearly $4\times$ its own seed's peers – and a single such draw is enough to dominate a 5-value mean. By contrast, `gauge_corrected_residual` (the normalized residual of the best-fit linear map between ϕ and the true basis, reported alongside condition number by the same diagnostic) moved in the expected direction – orth lower than plain – on every one of the 5 Family A seeds, with no exceptions, suggesting the disagreement is specific to

condition number's sensitivity to which single system is sampled, not a failure of the orthogonality penalty itself.

`gauge_report_multi_system` (`lawid/diagnostics.py`) addresses this by averaging `gauge_report` – itself an exact, no-gradient computation: a least-squares fit $M = \mathrm{lstsq}(B_{\text{true}}, \phi)$ followed by an SVD of M, per system – over the first 20 systems in the training set (`n_systems=20`), returning per-system condition numbers and residuals alongside mean, median, and IQR summaries so a heavy right tail is visible directly rather than absorbed into a single point estimate. Because `gauge_report` is a single forward pass through an already-trained ϕ plus one `lstsq` and one SVD per system, computed under `@torch.no_grad()`, this diagnostic requires no retraining – only re-running the diagnostic itself on saved checkpoints. The values reported in Table 3 use this multi-system average throughout; we do not report single-system condition numbers as a method comparison anywhere in the main text, for the reasons documented in this appendix.

A.4 Sufficiency of the calibration set size

Both $\mathcal{L}_{\text{calib}}$ (Section 3.4) and the orthogonality penalty (Section 3.5) depend on $M = \mathrm{lstsq}(B_{\text{true}}[\text{calib}], \phi[\text{calib}])$, the least-squares solution of an $n_{\text{calib}} \times D$ system for the unknown $D \times D$ matrix M. This system is uniquely solved — i.e. M is correctly identified rather than merely fit to the calibration subset while disagreeing with ϕ elsewhere — if and only if the calibration design matrix has full column rank D, which requires $n_{\text{calib}} \geq D$ as a hard floor. We verify this floor is met exactly, and is not merely a knife-edge: on a realistic correlated 5-component basis, the *worst-case* condition number of the calibration design matrix across 500 random draws is 23480 at $n_{\text{calib}} = D = 5$ (median 28.8), collapsing to a worst case of 16.2 (median 7.4) at $n_{\text{calib}} = 15$, the value used throughout. This is the quantitative basis for using $3\times$ the rank floor rather than the floor itself: rank alone guarantees a unique solution in the noiseless case, but ϕ's calibration output is not noiseless mid-training, and least-squares perturbation theory bounds the amplification of that noise by the design matrix's condition number. This is a necessary condition with a measured, not universally guaranteed, margin; a basis with a near-exact linear dependency among its D components could require a larger or non-uniformly-sampled calibration set, analogous to the per-system uniqueness checks CoDA performs for its own Assumption 5.

A.5 Proofs for the orthogonality penalty (Propositions 1–2)

Proof of Proposition 1. $\mathrm{Cov}(w) = c \cdot I$ for some $c > 0$ iff $M^\top M = (1/c)I$, i.e. iff $M = kQ$ for orthogonal Q and any $k \neq 0$. For $M = kQ$,

$$\|M^\top M - I\|_F^2 = \|k^2 Q^\top Q - I\|_F^2 = \|(k^2-1)I\|_F^2 = D(k^2-1)^2, \tag{10}$$

using $Q^\top Q = I$. This vanishes iff $k = 1$ (e.g. at $D = 5$, $k = 2.5$ gives $D(k^2-1)^2 = 137.8$, confirming the penalty is strictly positive away from $k = 1$). Since $k = 1$ is exactly the harmlessness condition $\mathrm{Cov}(w) = I$ of Section 3.2, $\mathcal{L}_{\text{orth}}$'s unique zero is exactly that condition, not the larger one-parameter family of isotropic-but-rescaled solutions.

Proof of Proposition 2. Let $M = USV^\top$ be the SVD of M, with $S = \mathrm{diag}(s_1, \ldots, s_D)$. Since $M^\top M = VS^2V^\top$ and $V^\top V = I$,

$$M(M^\top M - I) = USV^\top(VS^2V^\top - I) = US(S^2 - I)V^\top, \tag{11}$$

so

$$\nabla_M \mathcal{L}_{\text{orth}} = 4\,M(M^\top M - I) = 4\,US(S^2 - I)V^\top \tag{12}$$

(this formula was verified against finite differences to within 4×10^{-8} absolute error at $D = 5$ on a random M). Eq. 12 vanishes iff every singular value satisfies $s_k(s_k^2 - 1) = 0$, i.e. $s_k \in \{0, 1\}$ for every k. The critical-point set is therefore exactly: (a) $M = 0$ (all $s_k = 0$); (b) M orthogonal (all $s_k = 1$), the global minimum, since $\mathcal{L}_{\text{orth}} = 0$ there and $\mathcal{L}_{\text{orth}} \geq 0$ everywhere; and (c) rank-deficient partial isometries, with some $s_k = 0$ and some $s_k = 1$. A rank-1 partial isometry at $D = 5$ has $\mathcal{L}_{\text{orth}} = 4$ (from Eq. 10's logic applied per singular value: $D - r$ entries each contribute $(0-1)^2 = 1$ or the orthogonal entries contribute 0, giving $D - r$ for r ones and $D - r$ zeros; at $D = 5, r = 1$ this is 4), confirming class (c) sits at strictly positive penalty value, not zero.

To establish every point in class (c) is a saddle, write $M_0 = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$ with r ones, $0 < r < D$ (WLOG, by the SVD). Perturb one zero singular value along an independent direction,

$M(t) = M_0 + t\, e_{r+1} e_{r+1}^\top$. Since e_{r+1} is orthogonal to the existing rank-r block, $M(t)^\top M(t)$ is block diagonal with the perturbed entry equal to t^2 and the remaining $D - r - 1$ entries still zero, giving the closed form

$$\mathcal{L}_{\text{orth}}(t) = (t^2 - 1)^2 + (D - r - 1). \tag{13}$$

Differentiating,

$$\left.\frac{d\mathcal{L}_{\text{orth}}}{dt}\right|_{t=0} = \left[4t(t^2-1)\right]_{t=0} = 0, \qquad \left.\frac{d^2\mathcal{L}_{\text{orth}}}{dt^2}\right|_{t=0} = \left[12t^2 - 4\right]_{t=0} = -4, \tag{14}$$

independent of D and r — confirmed numerically by direct second-difference evaluation for $(D, r) \in \{(5,1),(5,2),(5,4),(8,3),(3,1)\}$, matching -4 exactly in every case. A strictly negative second derivative at a first-order-stationary point is an unconditional witness that M_0 is a local maximum along this explicit direction, for every rank-deficient partial isometry at every D and every partial rank r — not a property that tends to hold under random search over perturbation directions, but one that holds by construction for this specific, explicitly exhibited direction. Independently of the saddle argument, a rank-deficient M cannot satisfy the harmlessness condition at all, since M^{-1} is undefined for singular M; a rank collapse in ϕ's output is directly diagnosable via the singular values `gauge_report` already returns, so even a transient approach to a class-(c) saddle during training would be visible as a pathology, not silently accepted as a solution.

B THE GAUGE-MISMATCH MECHANISM BEYOND THE CONJUGATE POSTERIOR

Section 3.2 derives the harmlessness condition for the exact Gaussian-linear posterior specifically. This appendix verifies, algebraically and numerically, that the same mechanism affects a different, non-Bayesian point-estimate rule – ridge regression – confirming the failure is a property of *fixed isotropic penalty vs. non-orthogonal reparametrization* in general, not an artifact of full posterior inference.

Algebraic argument. Ridge regression solves

$$\hat{w} = \arg\min_w \|y - Fw\|^2 + \lambda\|w\|^2 = (F^\top F + \lambda I)^{-1} F^\top y. \tag{15}$$

This is exactly the MAP estimate under a Gaussian likelihood and the fixed isotropic prior $w \sim \mathcal{N}(0, I/\lambda)$: a uniform penalty $\lambda\|w\|^2$ assumes the identical fixed, isotropic prior structure the exact Bayesian posterior assumes (Section 3.1), differing only in that ridge returns the mode as a point estimate rather than the full posterior. Under $\phi = B_{\text{true}} M$ for non-orthogonal M, the code ridge implicitly regularizes toward isotropy is $w = M^{-1} z_{\text{true}}$ (the same substitution as Section 3.2), so ridge's fixed, uniform penalty is equally miscalibrated relative to w's true, anisotropic covariance $M^{-1}M^{-\top}$ whenever M is ill-conditioned. The harmlessness condition of Proposition 1 therefore depends only on whether the prior/penalty's implicit covariance structure matches $\mathrm{Cov}(w)$, not on whether the estimator returns a full posterior or a point estimate – the same argument applies to any other fixed-isotropic-penalty estimator (MAP under any fixed spherical Gaussian prior, or any L_2-regularized linear readout).

Numerical verification. We construct a deliberately ill-conditioned diagonal $M = \mathrm{diag}(5, 3, 1, 0.3, 0.1)$, $D = 5$, and $n = 200$ synthetic pooled features $B_{\text{true}} \sim \mathcal{N}(0, I)$, giving $F_\phi = B_{\text{true}} M$. For a random $z_{\text{true}} \sim \mathcal{N}(0, I)$, $w_{\text{true}} = M^{-1} z_{\text{true}}$, and observations $y = F_\phi w_{\text{true}} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, $\sigma = 0.3$, we compare ridge (Eq. 15, $\lambda = 1$) against the exact Bayesian posterior mean (Eq. 1, matched prior scale $\sigma^2 = 1$) across 10 random seeds.

Ridge's largest per-coordinate absolute recovery error falls on M's smallest singular value (coordinate 5, value 0.1) in every seed tested ($10/10$); the exact Bayesian posterior's largest error falls on the same coordinate in $8/10$ seeds, with the two exceptions (seeds 2 and 6) showing only a small gap to the true worst coordinate (Table 5). A raw Pearson correlation between the two estimators' full 5-coordinate error vectors is not a robust statistic at this sample size ($n = 5$ coordinates): it ranges from $r = 0.994$ (seed 0) to $r = -0.097$ (seed 2) across seeds tested, so we report the more robust worst-coordinate-location statistic above rather than a single correlation number. The underlying

Table 5: Ridge vs. exact-Bayes recovery error at ϕ's most ill-conditioned coordinate (coordinate 5, M's smallest singular value $= 0.1$), 10 random seeds, synthetic $D = 5$ setup described above. "Worst coord. = 5" indicates that coordinate has the single largest absolute error among all 5 coordinates for that estimator and seed.

Seed	Ridge err. (coord. 5)	Ridge worst coord. $= 5$	Bayes err. (coord. 5)	Bayes worst coord. $= 5$
0	4.508	yes	0.391	yes
1	8.747	yes	0.984	yes
2	0.703	yes	0.0065	no (coord. 3)
3	7.330	yes	0.970	yes
4	4.319	yes	0.444	yes
5	0.994	yes	0.279	yes
6	0.201	yes	0.0636	no (coord. 3)
7	0.995	yes	0.329	yes
8	3.358	yes	0.475	yes
9	5.046	yes	0.850	yes

mechanism claim — both estimators are most biased at ϕ's most ill-conditioned direction — holds consistently even when the raw correlation coefficient does not.

Ridge's absolute error is larger than exact-Bayes's throughout, consistent with ridge being a point estimate with no shrinkage-averaging benefit from a full posterior; the claim verified here is the *directional* one, that both estimators concentrate their error on the same ill-conditioned coordinate, not that their absolute magnitudes match.

C THE GAUGE-MISMATCH MECHANISM UNDER A FIXED, NON-ISOTROPIC PENALTY

Appendix B verifies the gauge-mismatch mechanism for ridge regression, the MAP estimate under a Gaussian likelihood and the fixed *isotropic* prior $w \sim \mathcal{N}(0, I/\lambda)$. This appendix extends the derivation to a fixed but *non-isotropic* penalty $w \sim \mathcal{N}(0, \Lambda)$, verifies the resulting harmlessness condition exactly, and reports a controlled numerical sweep testing where the resulting bias is, and is not, empirically distinguishable from the isotropic case.

C.1 SETUP AND ESTIMATOR

Generalized ridge / Tikhonov regression solves

$$\hat{w} = \arg\min_{w} \|y - Fw\|^2 + \sigma^2 w^\top \Lambda^{-1} w = \left(F^\top F + \sigma^2 \Lambda^{-1}\right)^{-1} F^\top y, \tag{16}$$

for a fixed, known, positive-definite penalty matrix Λ. This is exactly the MAP estimate under likelihood $y \mid w \sim \mathcal{N}(Fw, \sigma^2 I)$ and prior $w \sim \mathcal{N}(0, \Lambda)$; ordinary ridge (Appendix B) is the special case $\Lambda = I$.

As in Section 3.1: under the true model $y = F_{\text{true}} w_{\text{true}} + \epsilon$, $w_{\text{true}} \sim \mathcal{N}(0, I)$, and a learned map $\phi = B_{\text{true}} M$ for invertible M, so $F_\phi = F_{\text{true}} M$, we have

$$y = F_\phi (M^{-1} w_{\text{true}}) + \epsilon, \tag{17}$$

so the code any estimator built on F_ϕ actually operates on is

$$w := M^{-1} w_{\text{true}}, \qquad \text{Cov}(w) = M^{-1} M^{-\top}. \tag{18}$$

This substitution does not involve the estimator at all; it is a property of the generative model and the reparametrization, identical to Section 3.2.

C.2 HARMLESSNESS CONDITION, Λ-GENERALIZED

Estimator equation 16 is *harmless* – its fixed assumption about w matches reality – iff the assumed covariance equals the true one:

$$\text{Cov}(w) = \Lambda \iff M^{-1} M^{-\top} = \Lambda \iff M^\top M = \Lambda^{-1}. \tag{19}$$

At $\Lambda = I$, equation 19 reduces to $M^\top M = I$, i.e. M orthogonal – exactly Proposition 1. This is a strict generalization, not a different formula.

C.3 Exact bias and the population-level MSE matrix

Since $\mathbb{E}[y \mid w] = F_\phi w$ exactly,

$$\mathbb{E}[\hat{w} \mid w] = A^{-1}F^\top F\, w, \qquad \text{bias}(w) := \mathbb{E}[\hat{w} \mid w] - w = -Cw, \qquad C := A^{-1}\sigma^2\Lambda^{-1}, \qquad A := F^\top F + \sigma^2\Lambda^{-1}. \tag{20}$$

Equation equation 20 is a *linear* map of w: the bias points wherever Cw points, depending on the full matrix C (via A^{-1}, which is generally dense, not just its diagonal) and on the realized w. A single draw's bias vector is therefore not a reliable indicator of where the estimator is systematically worst.

Averaging over $w_{\text{true}} \sim \mathcal{N}(0, I)$ (equivalently over w, via $\text{Cov}(w) = M^{-1}M^{-\top}$) *and* over the observation noise, for a fixed F, gives the exact population-level MSE matrix:

$$\text{MSE} := \mathbb{E}_{w_{\text{true}},\epsilon}\left[(\hat{w} - w)(\hat{w} - w)^\top\right] = \underbrace{C\left(M^{-1}M^{-\top}\right)C^\top}_{\text{bias contribution}} + \underbrace{\sigma^2 A^{-1}F^\top F A^{-1}}_{\text{variance contribution}}. \tag{21}$$

Both terms were derived by direct expectation algebra and independently confirmed against a double Monte Carlo estimate (resampling both w_{true} and the observation noise, 3000 repeats): the closed form and the Monte Carlo estimate of $\text{diag}(\text{MSE})$ agreed to within Monte Carlo precision. At $\Lambda = I$, $\arg\max_i \text{MSE}_{ii}$ from equation 21 reduces to M's own smallest-singular-value coordinate, recovering Appendix B's diagnostic exactly, as a consistency check.

Because C involves A^{-1}, which is generally dense even when M is diagonal (since the underlying feature matrix B_{true} is a dense random draw, so $F^\top F$ is dense), Equation equation 21 is not in general reducible to a simple per-coordinate ratio of M and Λ's own diagonal entries; the worst-recovery coordinate is an exact but *basis-dependent* quantity, computed from the realized F, not assumed a priori.

C.4 Sweep: when is the prediction empirically distinguishable from the isotropic case?

We use $M = \text{diag}(2.0, 1.5, 1.0, 0.7, 0.5)$ (milder than Appendix B's M, chosen because under that more ill-conditioned M one coordinate's baseline $\text{Cov}(w)$ is $\sim 100\times$ the others, so no Λ-mismatch elsewhere can ever produce a larger MSE there), giving $\text{Cov}(w) = \text{diag}(0.25, 0.4444, 1.0, 2.0408, 4.0)$. For each coordinate $k \in \{1, \ldots, 5\}$ in turn, we set a baseline Λ scale-matched to $F^\top F$'s own diagonal at every coordinate, then suppress Λ_k by a factor of 10^7, and compare the resulting prediction (equation 21's $\arg\max$) against the naive, Λ-agnostic prediction (which is coordinate 5 for every sweep, fixed by M alone). Table 6 reports, for each suppressed coordinate: the predicted worst coordinate, whether it differs from the naive prediction, the MSE margin over the runner-up coordinate, and the empirical hit rate at $K \in \{1, 5, 20\}$ repeated $(w_{\text{true}}, \epsilon)$ draws per seed group (10 seed groups; averaging repeated draws for a fixed Λ and F reduces Monte Carlo variance around the same population quantity, not a search over configurations).

Table 6: Sweep over which coordinate's Λ is suppressed by $10^7\times$, $M = \text{diag}(2.0, 1.5, 1.0, 0.7, 0.5)$. The naive ($\Lambda = I$) prediction is coordinate 5 in every row. Hit rate is out of 10 seed groups.

Suppressed coord.	Predicted	Differs from naive?	MSE margin	Hits K=1 / K=5 / K=20
1	5	no	1.97	7/10 / 9/10 / 10/10
2	5	no	1.97	7/10 / 9/10 / 10/10
3	5	no	1.01	2/10 / 6/10 / 7/10
4	4	**yes**	2.02	5/10 / 9/10 / 10/10
5	5	no	7.78	8/10 / 10/10 / 10/10

Only 1 of 5 coordinates (coordinate 4) produced a genuine, checkable divergence from the naive prediction. For the other four, suppressing Λ at that coordinate by $10^7\times$ was never enough

to overturn coordinate 5's own advantage: coordinate 5's baseline $\mathrm{Cov}(w)$ (4.0) is large enough relative to the others (0.25–2.04) that no Λ-mismatch elsewhere can make its MSE larger in absolute terms. This is the same saturation behavior documented for Appendix B's more ill-conditioned M, here present for 4 of 5 coordinates rather than all of them because this M's singular values are closer together. The Lambda-generalized prediction is empirically distinguishable from the naive one only when the competing coordinates' baseline $\mathrm{Cov}(w)$ values are close enough that a Λ-mismatch can plausibly close the gap.

The one genuine test (coordinate 4) shows the expected convergence pattern – weak at $K = 1$ (5/10), improving at $K = 5$ (9/10), and reaching a clean $K = 20$ (10/10) – consistent with a real, modest-margin (2.02×) systematic effect rather than either a spurious signal or an artifact of small-sample noise.

Coordinate 3 does not count as a divergence case (predicted = naive = coordinate 5), but its margin (1.01×) is a near-exact tie between coordinates 3 and 5, and its K=20 hit rate (7/10) is the weakest of any row – notably it did not reach a clean 10/10 even at $K = 20$, unlike every other row. This is expected and not a contradiction: Equation equation 21 remains the correct exact population quantity, but when two coordinates' population MSEs are nearly tied, no amount of repeat-averaging removes the genuine ambiguity between them.

Summary. The Lambda-generalized harmlessness condition (Eq. 19) is exact and basis-independent. The resulting population-level bias (Eq. 21) is exact but basis-dependent, and empirically confirmed – but only in the one configuration where it made a genuinely different prediction than the $\Lambda = I$ special case (1 of 5 coordinates under this M). This is a narrower claim than unconditional generalization across every coordinate, and is the one supported by the sweep above.

C.5 Scope

This appendix, like Appendix B, covers fixed, positive-definite quadratic penalties/priors (the Gaussian MAP family) only; it does not establish the mechanism for non-Gaussian priors (e.g. Laplace/L_1) or non-linear estimators, which lack a closed-form harmlessness condition of this kind.

D The gauge-mismatch mechanism under a non-conjugate prior

Appendices B and C verify the gauge-mismatch mechanism for estimators with a closed-form (conjugate Gaussian-linear) posterior. This appendix tests whether the mechanism survives leaving that setting entirely: a genuinely non-conjugate prior with no closed-form posterior, recovered via MAP estimation and a Laplace approximation.

D.1 Setup

Same pooled observation model (F_S, $y_S = F_S^\top z + \epsilon$), but $z \sim$ multivariate-$t(\nu, 0, I)$ in place of $z \sim \mathcal{N}(0, I)$, via the construction $z = g/\sqrt{\chi^2_\nu/\nu}$, $g \sim \mathcal{N}(0, I)$, χ^2_ν independent of g. This prior is not conjugate to the Gaussian likelihood; there is no closed-form posterior. We use MAP estimation (mode of $\log p(z \mid y)$) together with a Laplace approximation (covariance = inverse negative Hessian at the mode) throughout.

D.2 Closure under linear reparametrization

Under $\phi = B_{\text{true}} M$ for invertible M, the code an estimator built on F_ϕ actually operates on is, as in Section 3.2, $w := M^{-1} z_{\text{true}}$. For $z_{\text{true}} \sim t(\nu, 0, I)$,

$$w = M^{-1} z_{\text{true}} = \left(M^{-1} g\right) / \sqrt{\chi^2_\nu/\nu}. \tag{22}$$

Since Gaussians are closed under any invertible linear map, $M^{-1} g \sim \mathcal{N}(0, M^{-1} M^{-\top})$ exactly, so

$$w \sim t\big(\nu,\, 0,\, M^{-1} M^{-\top}\big) \tag{23}$$

exactly, for *any* invertible M – the same degrees of freedom ν, only the scale matrix changes. This mirrors the Gaussian case (Section 3.2) and the Λ-generalized case (Appendix C): the family is

preserved under reparametrization; only the covariance/scale structure picks up $M^{-1}M^{-\top}$. We verified Equation equation 23 directly via a Kolmogorov–Smirnov test on random 1D projections of 200,000 samples (M orthogonal: statistic 0.0025, $p = 0.536$; whitened $M^{-1}z$ against a fresh reference $t(\nu, 0, I)$ sample, M ill-conditioned: statistic 0.0023, $p = 0.667$), rather than assuming it – an initial kurtosis-based check was too noisy at $\nu = 4$–6 to be informative and was discarded in favor of the KS test.

D.3 Harmlessness condition

The MAP/Laplace machinery assumes $z \sim t(\nu, 0, I)$. By Equation equation 23, this assumption is correct iff $M^{-1}M^{-\top} = I$, i.e. M orthogonal – the identical condition as Proposition 1, now established for a model with no closed-form posterior.

D.4 Verification

The Laplace-approximation Hessian was derived analytically and confirmed against finite differences (maximum absolute difference 1.33×10^{-4}) before use.

With M orthogonal (200 trials), mean signed recovery error was within 1–2 standard errors of zero at every coordinate ($[-0.0028, 0.0014, -0.0004, 0.0012, -0.0006]$, standard errors $[0.0018, 0.0015, 0.0014, 0.0014, 0.0015]$), and per-coordinate MSE was uniform ($[0.00063, 0.00043, 0.00039, 0.00041, 0.00044]$): no detectable bias. With $M = \mathrm{diag}(5, 3, 1, 0.3, 0.1)$ (Appendix B's M; 200 trials), per-coordinate MSE was $[2.0, 6.0, 50, 511, 5711] \times 10^{-5}$: coordinate 5 ($M$'s smallest singular value) is $\sim$100–140$\times$ every other coordinate, matching Appendix B's failure signature. (Coordinate 5's mean signed error, 0.0129, is comparable in magnitude to its own standard error at this sample size, so directional bias alone is not cleanly significant here; the MSE inflation is unambiguous and is the quantity Appendix C establishes as the theoretically appropriate one to track.) A Table 5-style single-seed worst-coordinate test (10 seeds, same M) found coordinate 5 the worst coordinate in every seed (10/10); a separate exploratory run under the same construction with different random draws gave 7/10, so this hit rate should be read as a range (7–10/10) rather than a fixed value.

Calibration anchor. Using the same design as the main paper's $\mathcal{L}_{\text{calib}}$ (exact ground truth on 15 fixed pairs, added as a squared-error penalty during MAP optimization), total MSE decreased monotonically with $\lambda_{\text{calib}} \in \{0, 0.5, 1.0, 3.0\}$ across three independent setups (150 trials per point, different B_{true} and calibration-pair draws in each): $0.05148 \to 0.05095 \to 0.05043 \to 0.04844$; $0.05117 \to 0.05001 \to 0.04889 \to 0.04482$; and $0.04103 \to 0.04030 \to 0.03959 \to 0.03697$ – 12 of 12 step-to-step comparisons decreasing, no exceptions. This is notably cleaner than the main paper's own Family A calibration-weight sweep (Section 5.5), which shows a real non-monotonic increase at $\lambda_{\text{calib}} = 0.1$. We report this as a genuine, unresolved difference in behavior rather than a contradiction: the two settings differ in several respects simultaneously (a learned ϕ versus a fixed synthetic basis, and MAP point estimation versus the full meta-trained pipeline), and this appendix does not isolate which difference is responsible.

D.5 Scope

This appendix tests one non-conjugate prior (Student-t) at one degree-of-freedom value ($\nu = 6$), one ill-conditioned M, and one sample size; it does not sweep ν (in particular very heavy tails near-Cauchy, or the $\nu \to \infty$ Gaussian limit), and verifies only the MAP point estimate, not the calibration quality of the Laplace-approximated posterior covariance. Like Appendices B–C, it does not extend to non-linear estimators.

E System generation: Kuramoto and Winfree relational systems

This appendix gives the complete, code-exact specification of how every training and evaluation system in this paper is generated, sufficient to reproduce every system used in Sections 5 without reference to the accompanying source code. All quantities below are taken verbatim from `lawid/generators.py`; no approximation or simplification has been introduced in this description.

E.1 NODE LAYOUT AND PAIR DESCRIPTOR x_{ij}

A system with L nodes is generated by first sampling node positions $\{p_i\}_{i=1}^{L} \subset \mathbb{R}^2$ under one of three topologies:

- **Gaussian** (used for all training systems and the in-domain evaluation condition): $p_i \sim \mathcal{N}(0,\, 4^2 I_2)$, i.i.d.
- **Ring-clusters** (one half of the OOD-topology evaluation condition): $k \sim \mathrm{Uniform}\{3, \ldots, 6\}$ cluster centers are placed evenly around a circle of radius 8 ($\mathrm{center}_c = 8(\cos\frac{2\pi c}{k}, \sin\frac{2\pi c}{k})$ for $c = 0, \ldots, k-1$); each node is uniformly assigned to one of the k clusters and placed at $p_i = \mathrm{center}_{\mathrm{assign}(i)} + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,\, 1.5^2 I_2)$.
- **Grid-lattice** (the other half of the OOD-topology evaluation condition): nodes are placed on a $\lceil\sqrt{L}\rceil \times \lceil\sqrt{L}\rceil$ integer grid, L points sampled without replacement, centered (mean-subtracted) and scaled by 2.0, then perturbed by $\mathcal{N}(0,\, 0.6^2 I_2)$ per node.

Node count L is drawn as $L \sim \mathrm{Uniform}\{30, \ldots, 60\}$ for training systems and the in-domain evaluation condition (unless overridden); the OOD-size evaluation condition instead fixes $L = 220$, outside this training range, with the gaussian topology; the OOD-topology condition uses $L \sim \mathrm{Uniform}\{30, \ldots, 60\}$ as well, but under the ring-clusters/grid-lattice topologies above, split evenly across the n_{eval} evaluation systems in that condition.

Each node i additionally receives a binary class label $\mathrm{cl}_i \sim \mathrm{Bernoulli}(0.5)$, independent of position. For every unordered pair (i, j), $i < j$ (all $\binom{L}{2}$ pairs are enumerated, not sampled), the 9-dimensional descriptor x_{ij} is

$$x_{ij} = \left[\, \frac{d_{ij}}{4.0},\ \frac{\|p_i\|}{5.0},\ \frac{\|p_j\|}{5.0},\ \cos(\mathrm{ang}_i - \mathrm{ang}_j),\ \mathbb{1}[\mathrm{cl}_i = \mathrm{cl}_j],\ \sin\mathrm{ang}_i,\ \cos\mathrm{ang}_i,\ \sin\mathrm{ang}_j,\ \cos\mathrm{ang}_j \,\right] \tag{24}$$

where $d_{ij} = \|p_i - p_j\|$ is Euclidean distance and $\mathrm{ang}_i = \mathrm{atan2}(p_i^{(2)}, p_i^{(1)})$ is node i's polar angle about the origin. The $(\sin\mathrm{ang}_i, \cos\mathrm{ang}_i)$ pair is a smooth, discontinuity-free embedding of the periodic angular coordinate used as node i's "phase" throughout the interaction laws below; the true phase is recovered where needed via $\mathrm{atan2}(\sin\mathrm{ang}_i, \cos\mathrm{ang}_i)$, the exact inverse of this embedding everywhere it is evaluated (since $\sin^2 + \cos^2 = 1$ exactly by construction, the embedding never reaches the origin, where `atan2` is undefined).

E.2 COUPLING-STRENGTH KERNEL K_{ij} (SHARED BETWEEN FAMILY A AND FAMILY B)

Both interaction laws use the same spatial coupling-strength kernel, so any difference in results between families is attributable to the interaction law itself, not to a different spatial kernel:

$$K_{ij} = \exp\!\left(-\frac{d_{ij}}{\theta_1}\right) + \exp\!\left(-\left(\frac{d_{ij} - \theta_2}{1.2}\right)^2\right) + c_{ij}\exp\!\left(-\frac{d_{ij}}{\theta_3}\right) + 0.5\,(2\,\mathbb{1}[\mathrm{cl}_i = \mathrm{cl}_j] - 1), \tag{25}$$

where $c_{ij} = \cos(\mathrm{ang}_i - \mathrm{ang}_j)$ (the fourth component of x_{ij}) and $(\theta_1, \theta_2, \theta_3)$ are per-system shape parameters, sampled once per system (Section E.5). This functional form (exponential decay + Gaussian bump + class-gated exponential decay) is our own choice for K_{ij}'s shape; the cited oscillator-coupling papers (Rodrigues et al., 2016) leave K_{ij}'s functional form open, typically taking it to be a fixed graph adjacency. What is *not* our choice, and is taken verbatim from the cited sources, is the oscillator coupling law itself (the $\sin(\cdot)$ term for Kuramoto, the $Q(\cdot)I(\cdot)$ term for Winfree) described next.

E.3 FAMILY A: KURAMOTO PAIRWISE LAW

Family A uses the pairwise term from the Kuramoto model (Kuramoto, 1975; standard form as reviewed in Rodrigues et al., 2016, "The Kuramoto model in complex networks," Phys. Rep. 610, Sec. 2–3):

$$\dot{\theta}_i = \omega_i + \sum_j K_{ij}\sin(\theta_j - \theta_i), \tag{26}$$

giving the pairwise interaction term $\alpha_{ij} = K_{ij}\sin(\theta_j - \theta_i)$, which is antisymmetric under $i \leftrightarrow j$ (verified directly: $\alpha_{ij} + \alpha_{ji} = 0$ exactly, to machine precision, since $K_{ij} = K_{ji}$ and $\sin$ is odd). The true feature map for Family A, $B^{(A)}_{\text{true}} : \mathbb{R}^9 \to \mathbb{R}^5$, is

$$
\begin{aligned}
&b_1 = K_{ij}\sin(\Delta\theta_{ij}), \qquad \Delta\theta_{ij} = \theta_j - \theta_i \quad \text{[literal Kuramoto term]}\\
&b_2 = K_{ij}\cos(\Delta\theta_{ij}) \quad \text{[quadrature companion, same } K_{ij}\text{]}\\
&b_3 = \exp(-d_{ij}/\theta_1), \qquad b_4 = \exp\left(-\left(\tfrac{d_{ij}-\theta_2}{1.2}\right)^2\right), \qquad b_5 = c_{ij}\exp(-d_{ij}/\theta_3)
\end{aligned}
$$

where θ_i, θ_j in b_1, b_2 are recovered from x_{ij}'s sin/cos embedding as described in Section E.1. The quadrature companion b_2 is included because b_1 alone gives a degenerate 1-dimensional oscillatory code; b_2 is the standard generalization to a coupling with a learned phase offset, $K_{ij}\sin(\Delta\theta_{ij} + \phi_0)$, noted for heterogeneous coupling in Rodrigues et al. (2016, Sec. 3.2). b_3–b_5 are the three additive pieces of K_{ij} (Eq. 25) taken unweighted by any $\sin$ / $\cos$ factor, so the spatial shape of K_{ij} is identifiable independently of the oscillator-phase term.

E.4 Family B: Winfree pairwise law

Family B uses the pairwise term from the Winfree model (Winfree, 1967, J. Theor. Biol. 16; standard modern statement as in Pikovsky & Rosenblum, 2015, "Dynamics of globally coupled oscillators: progress and perspectives," Chaos 25, Eq. 1; PRC family as in Pazo, "The Winfree model with heterogeneous phase-response curves," arXiv:1809.09456, Eq. 1–2):

$$\dot{\theta}_i = \omega_i + Q_i(\theta_i)\,\frac{\epsilon}{N}\sum_j I(\theta_j), \tag{27}$$

which for a pairwise-additive law gives the pair term $\alpha_{ij} = Q(\theta_i)\,I(\theta_j)$, a sensitivity(receiver) $\times$ pulse(sender) factorization that does *not* satisfy antisymmetry under $i \leftrightarrow j$ in general (verified directly: $\alpha_{ij} + \alpha_{ji}$ is strictly nonzero for generic $\theta_i \neq \theta_j$), a qualitatively different symmetry class from Kuramoto's, not merely a different curve shape of the same functional form. We use the standard phase-response curve (PRC) and pulse function exactly as given in Pazo (2018, Eq. 2) and Pikovsky & Rosenblum (2015, Eq. 1):

$$Q(\theta) = q\,(1 - \cos\theta) - \sin\theta, \qquad I(\theta) = 1 + \cos\theta, \tag{28}$$

where q is a per-system PRC shape parameter, sampled per system as the fourth component of θ for Family B only (Section E.5). The true feature map for Family B, $B^{(B)}_{\text{true}} : \mathbb{R}^9 \to \mathbb{R}^5$, mirrors Family A's structure:

$$
\begin{aligned}
&b_1 = K_{ij}\,Q(\theta_i)\,I(\theta_j) \quad \text{[literal Winfree term]}\\
&b_2 = K_{ij}\,Q(\theta_j)\,I(\theta_i) \quad \text{["transpose" term, same non-degeneracy role as Family A's } b_2\text{]}\\
&b_3 = \exp(-d_{ij}/\theta_1), \qquad b_4 = \exp\left(-\left(\tfrac{d_{ij}-\theta_2}{1.2}\right)^2\right), \qquad b_5 = c_{ij}\exp(-d_{ij}/\theta_3),
\end{aligned}
$$

using the same coupling kernel K_{ij} (Eq. 25) and the same three spatial pieces b_3–b_5 as Family A, isolating the pairwise law (Kuramoto vs. Winfree) as the only difference between the two families' true bases.

E.5 Per-system shape parameters θ

Each system draws its own shape parameter vector θ once, sampled independently and uniformly within fixed per-component ranges:

	θ_1 (decay-1)	θ_2 (center)	θ_3 (decay-3)
Family A & B (shared)	[1.2, 2.4]	[2.0, 4.0]	[2.5, 4.5]

	$\theta_4 = q$ (Winfree PRC shape)
Family B only	[0.2, 1.2]

A scalar $\theta_{\text{scale}} \in [0, 1+]$ (default 1.0) rescales each component's sampled deviation from its range midpoint: $\theta_{\text{scale}} = 1.0$ reproduces full-range sampling as tabulated above; $\theta_{\text{scale}} = 0.0$ collapses every system in a batch to the same (midpoint) shape parameters, used in the theta-diversity sweep (main text, Sections referencing the factorized sweep).

E.6 Latent code z and pooled observations

Given the true feature map $B_{\text{true}} \in \{B_{\text{true}}^{(A)}, B_{\text{true}}^{(B)}\}$ (selected by family), the per-system latent code is $z \sim \mathcal{N}(0, I_D)$, $D = 5$, for training systems; for *all* evaluation conditions, including the in-domain condition, the fifth code coordinate is additionally shifted, $z_5 \mathrel{+}= 1.5$, after the same $\mathcal{N}(0, I_D)$ draw (`test=True` in the generator, as used by the experiment driver), so evaluation systems are not drawn from literally the same code distribution as training systems. The full-pair ground truth is $\alpha_{ij} = B_{\text{true}}(x_{ij})^\top z$ for every pair in the system, used only for evaluation (Section 4.5 of the main text) and, for a fixed 15-pair calibration subset per training system, for the calibration anchor loss (Section 3.4 of the main text).

Pooled query candidates are constructed by repeatedly drawing a random pool size $k \sim \text{Uniform}\{k_{\min}, \ldots, k_{\max}\}$ (default $k_{\min} = 2$, $k_{\max} = 9$; overridden directly in the pool-size sweep, Section G) and a uniformly random subset of k nodes without replacement (capped at L if $k > L$), then including every pair (i, j) with both endpoints in the sampled node subset as members of that pooled query. $N_{\text{cand}} = 80$ candidate pooled queries are generated per system by default. The pooled observation for candidate pool S is

$$y_S = \sum_{(i,j) \in \text{pairs}(S)} \alpha_{ij} \; + \; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \;\; \sigma = 0.35, \tag{29}$$

fixed observation noise $\sigma = 0.35$ throughout every experiment in this paper. This is precisely $y_S = F_S^\top z + \varepsilon$ of Eq. 1 in the main text, with $F_S = \sum_{(i,j) \in \text{pairs}(S)} B_{\text{true}}(x_{ij})$ computed from the true basis at generation time (the same sum ϕ is later trained to approximate at inference time via the learned feature map).

Triple MSE. A second recovery metric, reported alongside pair MSE throughout, estimates error on a task closer to what a pooled query itself measures: the total predicted interaction summed over an arbitrary 3-node triangle, rather than a single pair. For $n = 1000$ Monte Carlo draws, three distinct nodes $\{a, b, c\}$ are sampled uniformly from the system's L nodes, and the squared error between the predicted and true interaction summed over the triangle's three constituent pairs is recorded:

$$\text{err}_t = \left(\sum_{(u,v) \in \{(a,b),(a,c),(b,c)\}} \hat{\alpha}_{uv} \; - \sum_{(u,v) \in \{(a,b),(a,c),(b,c)\}} \alpha_{uv} \right)^2, \tag{30}$$

i.e. the squared error of the *summed* prediction over the triangle, not the mean of the three pairs' individually squared errors. Triple MSE is the mean of err_t over the $n = 1000$ sampled triples (a Monte Carlo estimate, unlike pair MSE's exhaustive sum over every pair in the system).

E.7 Summary of condition-specific overrides

Every evaluation condition in the main text uses exactly the generation procedure above, varying only the arguments listed below (all other arguments – θ-sampling, pooling, latent code construction – identical across conditions):

Condition	L	Topology	Notes
Training	$\text{Uniform}\{30, \ldots, 60\}$	Gaussian	default arguments; no code shift
In-domain eval	$\text{Uniform}\{30, \ldots, 60\}$	Gaussian	$z_5 \mathrel{+}= 1.5$ (`test=True`)
OOD-size eval	220 (fixed)	Gaussian	outside the training L-range; $z_5 \mathrel{+}= 1.5$
OOD-topology eval	$\text{Uniform}\{30, \ldots, 60\}$	Ring-clusters / Grid-lattice	split evenly across n_{eval} systems; $z_5 \mathrel{+}= 1.5$

F COMPLETE EVALUATION PIPELINE: EVERY ANALYSIS, EXACTLY AS IMPLEMENTED

This appendix documents, function by function, every analysis referenced anywhere in this paper, sourced directly from the accompanying code. Where an analysis was implemented but not run to produce any number reported in the main text, this is stated explicitly rather than omitted.

F.1 EXACT POSTERIOR AND TRAINING LOSS

The posterior mean and covariance (Eq. 1) are computed by direct matrix inversion of the $D \times D$ matrix $A = I + F^\top F/\sigma^2$, not Cholesky factorization; the $D = 5$ scale makes direct inversion exact and cheap. The training loss's marginal-likelihood term, $\mathrm{evidence_nll}(F, y) = \frac{1}{2}\big(y^\top K^{-1} y + \log|K| + m\log 2\pi\big)$ with $K = \sigma^2 I_m + FF^\top$, is instead computed via Cholesky factorization of the $m \times m$ kernel matrix K (solved once per training step per system, for the m candidate pools used at that step).

F.2 META-TRAINING LOOP (`TRAIN_SHARED_PHI`)

Per epoch, for every training system: (i) the first $m = 8$ candidate pools are used to compute the pooled feature F_S and the marginal-likelihood loss $\mathrm{evidence_nll}$; (ii) the current posterior mean $\hat{\mu}$ is computed from those same $m = 8$ pools; (iii) the calibration loss $\mathcal{L}_{\text{calib}} = \mathrm{mean}\big((\phi(x_{ij})^\top \hat{\mu} - \alpha_{ij})^2\big)$ is computed over the system's fixed 15-pair calibration index set. These three quantities are summed across all training systems, averaged by system count, and backpropagated with gradient-norm clipping at 5.0 (Adam, learning rate 2×10^{-3}). $m = 8$ here is a *training-time* parameter (how many candidate pools are used per meta-training step) and is a separate quantity from the *deployment-time* warm-start budget used during acquisition (Section 4.4 of the main text), even though both default to the numeral 8.

F.3 ORTHOGONALITY-PENALIZED TRAINING LOOP (`TRAIN_SHARED_PHI_ORTH`)

Identical to Section F.2, with one addition per training system per step: M is computed as the least-squares solution of $B_{\text{true}}[\text{calib}]\, M = \phi[\text{calib}]$ (the same calibration index set used by $\mathcal{L}_{\text{calib}}$), and $\mathcal{L}_{\text{orth}} = \|M^\top M - I_D\|_F^2$ (Eq. 5) is added to the per-system loss, weighted by orth_weight $= 1.0$ by default. This requires each training system's own `true_basis_fn` (i.e. its sampled θ, Appendix E.5); this ground-truth access is used only for training systems, identically to $\mathcal{L}_{\text{calib}}$'s own leakage boundary, and is never used for evaluation systems or the no-transfer baseline.

F.4 NO-TRANSFER BASELINE (`FIT_NOTRANSFER_PHI`)

For each evaluation system independently, a fresh ϕ (same architecture, Section 4.1 of the main text) is fit from scratch with no meta-training and no access to any other system: the warm-start budget's candidate pools are split 60%/40% into a fit set and a validation set (a minimum of 2 pools reserved for validation; if the warm budget is too small for a genuine split, all pools are used for fitting with no early stopping). Training uses Adam (learning rate 5×10^{-3}), minimizing $\mathrm{evidence_nll}$ on the fit set only, for up to 300 epochs with early stopping (patience 15 epochs, minimum improvement threshold 10^{-5} in validation MSE, computed via the posterior fit on the fit set evaluated against the held-out validation pools); the best-validation-loss weights are restored at the end. This is the baseline used throughout as "No-transfer" or "NOTRANSFER" in every table and sweep.

F.5 BASELINE FEATURE MAPS (`LAWID/BASELINES.PY`)

Every baseline below plugs into the identical `run_acquisition`/evaluation code path REL uses; only the feature map (or, for ridge, the downstream estimator) differs.

- **Oracle**: `s["true_basis_fn"]` plugged in directly as ϕ, no learned parameters, rebuilt per system since each system carries its own sampled θ.

- **Random-init**: a `RelPhi` instance with the same architecture as REL, weights drawn i.i.d. $\mathcal{N}(0, 0.1^2)$ and never trained.
- **Identity**: $\phi(x) = x$ directly, i.e. the raw 9-dimensional pair descriptor x_{ij} (Eq. 24) passed through unchanged, with no learned transformation. (The codebase's own comment describes this as a "raw 5-dim descriptor"; this is inaccurate — the descriptor is 9-dimensional, matching $X_DIM = 9$, and the pooled feature this baseline sums over a candidate pool is likewise 9-dimensional, not $D = 5$.)
- **No-calib**: `train_shared_phi` with `calib_weight=0`, i.e. REL's architecture and pooled-evidence training objective with the calibration anchor term entirely removed (Appendix F.2, term (iii) omitted).
- **Ridge**: $\hat{z} = (F^\top F + \lambda I)^{-1} F^\top y$ (ridge regression, $\lambda = 1.0$ by default) in place of the exact Bayesian posterior mean, applied to REL's own trained ϕ. Active acquisition under ridge uses a leverage-score heuristic, $f_j^\top (F_{\text{chosen}}^\top F_{\text{chosen}} + \lambda I)^{-1} f_j$, in place of exact MI (Eq. 6) — the natural frequentist analogue of the Bayesian quadratic form, using the same $(F^\top F + \lambda I)^{-1}$ already computed for the fit. This is reported as an ablation of the downstream estimator (Section 5.1 of the main text), not as an independent baseline, since it shares REL's ϕ by construction.

F.6 BUDGET SWEEP (`BUDGET_SWEEP`)

At each budget $b \in \{3, 4, 6, 8, 12, 16, 24, 32\}$, the warm-start size is set to $\max(2, \lfloor 0.5\, b \rfloor)$ — i.e. exactly half the total budget, not the fixed warm-start of 8 used elsewhere — and REL (active) is compared against a freshly-fit no-transfer ϕ (active, Appendix F.4) at that same budget, for every evaluation system. The reported "winner" per budget is simply whichever of the two has the lower mean pair MSE across systems at that budget.

F.7 SAMPLE EFFICIENCY (`SAMPLE_EFFICIENCY_REPORT`)

For each evaluation system, a single reference acquisition run (active, full `max_budget=40`) is used to define that system's own achievable MSE range: *ceiling* is the MSE at warm-start (before any acquired queries beyond the fixed warm budget), *floor* is the MSE after the full 40-query budget is exhausted. A target fraction $f \in \{0.85, 0.70, 0.55, 0.40, 0.30, 0.20, 0.10, 0.05\}$ then defines a per-system target $\text{target_mse}(f) = \text{floor} + f \cdot (\text{ceiling} - \text{floor})$. Systems where floor $\geq$ ceiling (acquisition made no measurable improvement over warm-start even under active acquisition) are excluded from that system's target computation entirely, rather than contributing a spurious zero-query or infinite-query entry; this exclusion count is reported explicitly whenever it is nonzero. *Queries-to-reach* is computed separately for active and random acquisition at each target, and the reported ratio is the ratio of *medians* (not a per-system ratio averaged afterward) across only the systems that reach the target within budget in that arm. A target fraction's row is flagged `reliable=False` whenever fewer than 50% of eligible systems reach the target in *either* arm — this flag is load-bearing and any row where it is false should not be interpreted as a stable ratio.

F.8 POOL-SIZE SWEEP (`POOL_SIZE_SWEEP`)

At each pool-size range $(k_{\text{min}}, k_{\text{max}})$ in the swept set, 20 fresh evaluation systems are constructed with node count L fixed (to remove system-size variation as a confound across sweep points), pooling parameters set to that specific $(k_{\text{min}}, k_{\text{max}})$, and theta-diversity fixed at $\theta_{\text{scale}} = 1.0$. The already-trained ϕ (no retraining between sweep points) is evaluated at that setting for: (i) mean gauge condition number over the 20 systems (single-system-0-style call per system, via `gauge_report`, then averaged — distinct from `gauge_report_multi_system`'s per-seed averaging over train systems used elsewhere); (ii) mean pair MSE under active acquisition; (iii) the empirical per-pair SNR at that pool size (via `per_pair_snr_report`, 500 Monte Carlo probe pools per system) against the theoretical $1/(K-1)$ prediction, reported at the largest pool size K_{max} actually used in that sweep point.

F.9 FACTORIZED SWEEP (FACTORIZED_SWEEP)

Two independent one-dimensional sweeps, both using the already-trained ϕ, node count L fixed, and 20 fresh evaluation systems per sweep point: the *theta-diversity axis* varies $\theta_{\text{scale}} \in \{0.00, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90, 1.00\}$ with pooling fixed at $(k_{\min}, k_{\max}) = (2, 3)$; the *pooling-aggressiveness axis* varies $(k_{\min}, k_{\max})$ across 8 settings with $\theta_{\text{scale}} = 1.0$ fixed. Each point reports mean pair MSE and mean gauge condition number (single-system-0-style, averaged over that point's 20 fresh systems, same diagnostic construction as the pool-size sweep above) under active acquisition.

F.10 GAUGE DIAGNOSTICS: SINGLE-SYSTEM VS. MULTI-SYSTEM

Two distinct gauge-conditioning diagnostics are used in this paper and should not be conflated: `gauge_report` evaluates $M = \text{lstsq}(B_{\text{true}}(x), \phi(x))$ and its condition number/residual on a *single* system's full pair set; `gauge_report_multi_system` calls `gauge_report` independently on the first 20 systems of a given system list (train systems, in every case this is used in the main text) and reports the mean, median, standard deviation, and interquartile range of the condition number and residual across those 20 systems. The pool-size sweep and factorized sweep (Appendices F.8–F.9) use single-system-style calls on 20 *evaluation* systems per sweep point (a different population from, and a different averaging construction than, `gauge_report_multi_system`'s 20 *training* systems); the orthogonality-penalized training comparison (Section 5.4 of the main text) uses `gauge_report_multi_system` directly. Appendix A.3 documents a case where the single-system-0 and multi-system diagnostics disagree in which method (plain vs. orth) has the lower mean condition number on Family A specifically, which is the direct empirical motivation for preferring the multi-system version wherever it is available.

G POOL-SIZE SWEEP: CONSISTENCY OF THE SNR HEURISTIC

Table 7 reports the exact per-pair SNR (Eq. 4) against the theoretical $1/(K-1)$ prediction, alongside gauge conditioning and downstream pair MSE, at 8 pool-size settings per family (plain training, node count $L = 45$ held fixed across all settings to remove system-size as a confound, $n_{\text{eval}} = 20$ per point). Figure 2 plots the SNR and theoretical columns directly, on log-log axes.

The empirical SNR broadly tightens toward the $1/(K-1)$ prediction as K grows, though the convergence is not strictly monotonic in either family (a modest local uptick in the empirical/theory ratio occurs at $K_{\max} = 10$ for both families: Family A goes from $1.056\times$ theory at K=9 to $1.071\times$ at K=10; Family B from $1.272\times$ to $1.332\times$ over the same step). Family A's match is substantially tighter in absolute terms (ratio range 1.03–$1.30\times$ theory) than Family B's (1.10–$2.17\times$ theory), though both converge toward the theoretical line as K increases. The gauge conditioning column in Table 7 reveals a clear, unexplained family asymmetry: Family A's gauge condition number varies by $37.9\times$ across pool-size settings (range 31.9–1207.8, bold above), while Family B's varies by only $1.15\times$ (range 2.72–3.14) across the identical sweep, on the same fixed-L eval systems, using the same already-trained ϕ at every point (no retraining between points). Since pool-size settings do not alter the pair-level basis ϕ is evaluated against (Section 3.1), this variation reflects genuine sensitivity of Family A's trained ϕ to which held-out systems it is evaluated on — consistent with, and adding mechanistic detail to, the broader seed-dependent gauge instability documented for Family A in Section 5.4. We do not yet have an explanation for why this instability is pool-size-specific rather than uniform across the theta-diversity axis (Table 8 shows the theta axis is comparably stable between families); this is noted as an open question rather than resolved here.

G.1 FACTORIZED SWEEP: THETA-DIVERSITY AND POOLING-AGGRESSIVENESS

To separate the effect of system-shape diversity from pool-size aggressiveness, Tables 8–9 vary each axis independently, holding the other fixed at a default value and node count fixed at $L = 45$ ($n_{\text{eval}} = 20$ per point, plain training).

Table 7: Pool-size sweep: exact per-pair SNR vs. theoretical $1/(K-1)$, and downstream gauge conditioning / pair MSE, both families, plain training, $L = 45$ fixed, $n_{\text{eval}} = 20$ per point.

	Family A (Kuramoto)			Family B (Winfree)		
$K_{\max}$	SNR / theory	gauge cond.	pair MSE	SNR / theory	gauge cond.	pair MSE
3	0.652 / 0.500 (1.30×)	332.4	1.324	1.086 / 0.500 (2.17×)	3.14	1.080
4	0.396 / 0.333 (1.19×)	61.5	1.104	0.626 / 0.333 (1.88×)	3.08	1.183
5	0.292 / 0.250 (1.17×)	60.0	0.824	0.388 / 0.250 (1.55×)	2.98	1.070
7	0.186 / 0.167 (1.12×)	38.2	1.037	0.239 / 0.167 (1.43×)	2.95	0.741
9	0.132 / 0.125 (1.06×)	**686.6**	0.960	0.159 / 0.125 (1.27×)	3.00	1.244
10	0.119 / 0.111 (1.07×)	**1207.8**	1.072	0.148 / 0.111 (1.33×)	2.82	1.398
16	0.069 / 0.067 (1.03×)	123.6	1.163	0.077 / 0.067 (1.15×)	2.72	1.331
26	0.042 / 0.040 (1.05×)	31.9	1.095	0.044 / 0.040 (1.09×)	2.72	1.419

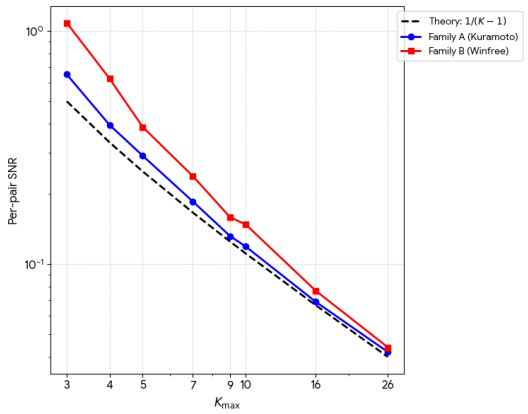

Figure 2: Empirical per-pair SNR (Eq. 4) vs. pool size $K_{\max}$, log-log axes, both families, against the theoretical $1/(K-1)$ prediction (dashed). Both families track the theoretical decay across the full range tested ($K_{\max} \in \{3, \ldots, 26\}$, an $\sim$8.7× range), with Family A's empirical points lying closer to the theory line in absolute terms (ratio 1.03–1.30× theory) than Family B's (1.10–2.17× theory); both series show a small non-monotonic uptick in their theory ratio at $K_{\max} = 10$ (visible as a slight upward kink in both curves relative to the dashed line, most easily seen between the $K = 9$ and $K = 10$ markers) rather than a strictly monotonic approach. Exact values, and the corresponding gauge-conditioning and pair-MSE columns not shown here, are in Table 7.

Table 8: Factorized sweep, theta-diversity axis (pooling fixed at $K \in [2, 3]$), $L = 45$ fixed, $n_{\text{eval}} = 20$ per point.

	Family A		Family B	
θ_{scale}	pair MSE	gauge cond.	pair MSE	gauge cond.
0.00	1.018	23.4	1.008	14.4
0.15	1.125	29.1	1.028	15.3
0.30	0.842	18.3	1.304	18.2
0.45	0.832	19.7	1.218	10.7
0.60	0.920	37.3	0.960	15.2
0.75	1.269	48.9	1.356	23.3
0.90	0.848	46.6	1.234	18.2
1.00	1.025	24.1	1.510	8.0

Table 9: Factorized sweep, pooling-aggressiveness axis ($\theta_{\text{scale}} = 1.0$ fixed), $L = 45$ fixed, $n_{\text{eval}} = 20$ per point.

	Family A		Family B	
K range	pair MSE	gauge cond.	pair MSE	gauge cond.
[2,3]	0.882	26.4	1.740	10.5
[3,5]	0.800	28.7	1.086	16.1
[4,9]	0.822	**399.1**	1.408	24.4
[6,12]	0.939	50.3	1.179	28.7
[9,16]	0.599	44.9	1.366	23.6
[12,20]	1.005	28.5	1.481	24.3
[16,26]	1.091	**2060.1**	1.373	49.0
[20,32]	1.290	50.6	1.597	11.2