THE COST OF EQUIVARIANCE IN NOISY LINEAR REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariance is a natural inductive bias for symmetric prediction tasks. Its benefit depends on where the symmetry is imposed. We ask what it costs to impose symmetry on a shared low-dimensional representation when receivers also have their own local features. In a paired Gaussian model, each receiver predicts its partner's coordinate from its own possibly noisy coordinate plus one common energy-constrained noisy linear broadcast; we minimize average risk over all fixed linear encoders and all measurable decoders. For up to half as many channels as input coordinates, we derive the exact minimum and characterize every optimal encoder Gram: equal-energy codes on selected coordinate pairs. On specified cycles and Boolean cubes, we also derive the exact optimum for translation-equivariant linear encoders with orthogonal latent actions. Their risk is strictly higher below the half-dimension threshold whenever local information is informative, so no decoder can recover the shortfall. Averaging optimal Grams over translations can exceed the rank budget; shared randomization instead restores symmetry in distribution. These exact floors separate information loss at the representation from approximation error at the decoder. In a prospectively specified 25-replicate CPU extension following five original replicates, frozen TabPFN-3 and TabICLv2 approach both floors; a separate ten-replicate cube study gives a second implementation check. An invertible post-noise transform preserves the floors while increasing ridge error, and both pretrained decoders retain low excess risk. The result is an exact cost of fixed encoder equivariance under matched budgets within this linear-Gaussian paired family.

1 INTRODUCTION

Equivariance is a natural inductive bias for symmetric prediction tasks. We study a shared encoder that compresses node features for per-node predictors, each of which also receives its own local feature. We show that imposing the task's translations on this bottleneck can strictly increase its Bayes risk, even when the decoder is unrestricted. This comparison optimizes the encoders under matched dimension, energy, and noise constraints.

We make this representation-design question exactly solvable. An even number of vertices carry independent Gaussian inputs. Each vertex wants its paired partner's input, sees its own possibly noisy input, and receives the same noisy linear broadcast: one shared representation whose width and total energy are the budget held fixed across the comparison. We minimize average prediction error across receivers, over every fixed linear encoder and every measurable decoder, so the resulting values are properties of the representation rather than of one fitted predictor. On the specified cycles and Boolean cubes, partners are antipodes and both the source law and the target respect translations. The encoder may be learned from earlier independent data but is fixed on fresh inputs. We compare all such linear encoders with translation-equivariant ones whose latent action is orthogonal, equivalently those with translation-invariant Grams.

Local features are central to the mechanism. A receiver can subtract its own contribution from a broadcast containing both partners, with attenuated subtraction when its observation is noisy. Concentrating the limited channels on a few selected pairs can therefore deliver more than spreading the same energy uniformly across vertices. Pair-sum coding is established analog network coding (Katti et al., 2007); the optimization question is whether any dense mixture can improve on it, and

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.

how much enforcing fixed symmetry costs. Gaussian computation broadcast supplies the general model and fixed-precoder decoder optimization (Tan & Jafar, 2026). Our contribution is the exact joint encoder optimum and symmetry comparison for this paired family.

Exact representation optima. For up to $n/2$ channels, we derive the global minimum and *every* optimal Gram: balanced rank-one codes on selected demand pairs, with equal energy. A sharp paired-entry inequality rules out better mixtures. We also derive the exact optimum under translation-invariant Gram constraints on the stated cycles and cubes. Below $n/2$ channels, and whenever local observations are informative, that invariant optimum is strictly worse than the unrestricted one. A sufficient neighborhood-volume certificate shows that the separation can persist when a receiver observes a local ball rather than a single coordinate.

The role of the representation budget. For the all-plus optimal pair-sum construction, averaging Grams over translations preserves energy but increases rank to $n/2$, which exceeds the channel budget in exactly the regime where the gap is strict. Random pair selection with shared, input-independent metadata preserves the unrestricted risk and restores symmetry in distribution. This distinguishes fixed encoder equivariance from distributional symmetry and explains why symmetrization can change the feasible implementation.

Separating information loss from decoder error. A conditional-risk identity transfers the representation floors to context-based predictors, including feature-only interactions among independent query rows. Declared receiver canonicalization makes representative evaluation valid without assuming that the predictor is rotation invariant. Independent complete realizations become supervised table rows: features are a receiver's permitted observations and context rows supply target labels. Frozen TabPFN-3 and TabICLv2 provide practical decoders without task-specific weight training; their labeled-context adaptation is not zero-shot prediction. A prospectively specified 25-replicate cycle extension and a separate ten-replicate cube study show low decoder excess relative to both floors. An invertible post-noise transform preserves information while increasing ridge error. Together with learned encoder structure and equal-floor controls, this gives a controlled test of what a representation retains and what a decoder can extract. The exact comparison concerns fixed linear encoders, Gaussian observations, and paired demands.

2 Related work

Linear coding with receiver side information. The Gaussian observation model is a specialization of wireless linear computation broadcast (Tan & Jafar, 2026). That work optimizes the decoder for a fixed precoder and the precoder for fixed decoders, with stationarity guarantees for alternating updates. We instead ask for the global optimum and the complete set of optimizers in one paired-demand family, which a stationary-point guarantee does not supply. Appendix C gives the parameter correspondence, including total-power normalization.

Earlier linear network coding already combines dimensionality reduction, power constraints, and estimation (Goela & Gastpar, 2012). Its fixed-decoder quadratic program and rank-relaxed lower bound do not directly solve the shared-encoder optimization here. Paired exchange also appears in Gaussian MIMO switching, whose inspected optimality theorems concern limiting signal-to-noise regimes (Wang et al., 2013). Side-information cancellation through a sum is established analog network coding (Katti et al., 2007); we use that construction rather than claim a new coding mechanism. Analog-source broadcast with receiver side information also predates this work (Piantanida et al., 2013), including fading MISO channels with feedback and high-SNR rate–distortion analysis. Here we ask for exact finite-noise values and all optimal Grams within the stated fixed-linear paired family. Gaussian complementary delivery has exact rate-distortion characterizations for long block codes (Timo et al., 2010). The finite-dimensional, fixed-linear encoder restriction in our results is therefore essential: these are representation-class optima, not unrestricted communication limits.

Noisy dimension reduction and joint source–channel coding. Schizas et al. (2007, Theorem 4) give an exact noisy single-sensor solution whose isotropic specialization recovers our no-local-information value; their noisy correlated-multisensor updates instead fix the other sensor maps. Floor et al. (2013, Sections 3–4) study cooperative and distributed Gaussian transmission to one common

receiver, with optimal linear power allocation and nonlinear pre-channel mappings. These are substantive precedents for the classical boundary case and encoding framework. Neither inspected result includes our receiver-specific private observations or fixed translation-invariant Gram comparison. Appendix C spells out the overlap and the distinct observation restrictions.

Classical inequalities and the specific optimization. Our proof uses standard Gaussian conditioning and Schur complements (Gallier, 2019). The trace comparison under pair pinching is an instance of Jensen's trace inequality (Hansen & Pedersen, 2003, Theorem 2.4). The remaining pairwise estimate and its equality conditions identify which mixtures can attain the resource bound. Equal power across nonzero singular directions alone would not establish localization on demand pairs. We center the contribution on the resulting exact values and complete optimal-Gram description, without claiming a new general conditioning or pinching principle.

Symmetry of representations and distributions. Probabilistic symmetry breaking and representations of equivariant conditional laws are established (Bloem-Reddy & Teh, 2020; Lawrence et al., 2025). Risk improvement by group averaging concerns the resulting predictor; it does not automatically preserve an implementation through a fixed noisy rank-constrained broadcast. In our example, averaging Grams of all-plus optimal pair-sum encoders preserves power but increases rank to $n/2$. Drawing one such Gram at random instead preserves feasibility per realization and restores equivariance in distribution, assuming shared selection metadata. Frame averaging and learned canonicalization already turn suitable backbone predictors into equivariant maps (Puny et al., 2022; Kaba et al., 2023); our receiver-coordinate wrapper is a specified instance of this established idea. Finite-sample symmetry benefits (Elesedy & Zaidi, 2021, Theorem 13) and symmetry-mismatch bias–variance tradeoffs (Huang et al., 2023, Theorem 2) compare fitted predictors. Our gap instead compares population Bayes risks under a common noisy representation budget, even when the task has the symmetry. We also distinguish active symmetries of a fixed graph from passive vertex relabeling (Huang et al., 2023): our acting groups are cycle rotations and cube XOR translations.

Frozen tabular decoders. TabPFN-3 (Grinsztajn et al., 2026) and TabICLv2 (Qu et al., 2026) supply pretrained regression decoders for the controlled observations. Their measured error indicates only proximity to representation-specific Bayes floors: it neither proves the encoder theorem nor bounds the models' internal computation. Ridge and inverse-transform controls retain the distinction between information retained by an observation and difficulty extracting that information.

3 Exact risk and the cost of a fixed symmetric bottleneck

Information interface. Let $n = 2m$ and let π be a fixed-point-free involution of the vertices; write $(Jx)_v = x_{\pi(v)}$. In each independent graph draw, take mutually independent

$$x \sim \mathcal{N}(0, I_n), \quad \zeta \sim \mathcal{N}(0, \sigma^2 I_n), \quad \eta \sim \mathcal{N}(0, \nu I_r), \qquad 0 \le \sigma^2 < \infty, \quad \nu > 0. \tag{1}$$

Receiver v predicts $Y_v = x_{\pi(v)}$ from $O_v = (u_v, z)$, where $u_v = x_v + \zeta_v$ and $z = Bx + \eta$. The same broadcast z is available to all receivers. The real encoder $B \in \mathbb{R}^{r \times n}$ is fixed independently of the current draw and satisfies $\|B\|_F^2 \le P$, equivalently $\mathbb{E}\|Bx\|^2 \le P$. Earlier independent training may choose B. Set $\beta = (1 + \sigma^2)^{-1}$ and $d = 1/\beta$. Let $D(B)$ be the infimum of the *total* MSE over all measurable receiver decoders, and let D_{free} be its infimum over feasible B. Per-vertex risks divide these quantities by n.

This is a special case of Gaussian linear computation broadcast (Tan & Jafar, 2026). Its Gaussian conditioning and side-information cancellation are established tools. Our optimization variable is a fixed linear encoder, not a general nonlinear or input-dependent encoder. Neither channel count nor P is a bit budget or a bound on computing cost. Gaussian *noise*, as well as the source, is essential for the stated MMSE formulas; specifying only its variance would suffice for linear prediction but not for unrestricted measurable decoding.

Posterior risk. Write $G = B^\top B$ and

$$Q = B^\top (BB^\top + \nu I_r)^{-1} B = G(G + \nu I_n)^{-1}, \qquad 0 \preceq Q \prec I_n. \tag{2}$$

Gaussian conditioning and conditional-mean orthogonality give

$$D(B) = n - \operatorname{tr} Q - \sum_{v} \frac{Q_{v,\pi(v)}^2}{d - Q_{vv}}. \tag{3}$$

Every denominator is positive, including with exact local observations. The trace is information from the broadcast alone; the final term is the additional gain from local side information. Optimizing the latter requires controlling where the entries of Q can concentrate, rather than its spectrum alone.

Lemma 1 (Paired-entry inequality and equality cases)**.** *For real symmetric $0 \preceq Q \prec I_n$, finite $d \geq 1$, and any fixed-point-free involution π,*

$$\sum_{v} \frac{Q_{v,\pi(v)}^2}{d - Q_{vv}} \leq \operatorname{tr}\big[Q^2(2dI_n - Q)^{-1}\big]. \tag{4}$$

Equality holds exactly when Q is block diagonal on the prescribed pairs and each pair block is zero or

$$\frac{\lambda}{2}\begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}, \qquad 0 < \lambda < 1, \quad \epsilon \in \{-1, 1\}. \tag{5}$$

The proof separates two scalar remainders, measuring diagonal imbalance and rank two within a pair, from an inverse-compression remainder measuring coupling between pairs. All three are nonnegative. The last step is a standard Schur-complement inverse inequality (Gallier, 2019); its zero-trace condition forces every cross-pair block to vanish. The complete argument, including zero blocks, is in Appendix A.

Theorem 2 (Unrestricted value and every optimal Gram)**.** *Under* (1)*, suppose $P > 0$ and $1 \leq r \leq n/2$. Then*

$$D_{\text{free}} = n - \frac{rP}{(1 - \beta/2)P + r\nu}. \tag{6}$$

Every optimal Gram consists of exactly r selected pair blocks

$$G = \frac{P}{2r} \bigoplus_{j=1}^{r} \begin{pmatrix} 1 & \epsilon_j \\ \epsilon_j & 1 \end{pmatrix} \oplus 0, \qquad \epsilon_j \in \{-1, 1\}, \tag{7}$$

after reordering the fixed pairs. Conversely every such Gram is optimal. All optimizers use full power, have rank r, and have all positive squared singular values equal to P/r.

For intuition, one attaining encoder sends $z_j = a(x_{u_j} + x_{\pi(u_j)}) + \eta_j$ on r pairs, with $a = \sqrt{P/(2r)}$. A selected receiver uses

$$\widehat{Y}_v = \frac{a}{(2 - \beta)a^2 + \nu}(z_j - \beta a u_v), \tag{8}$$

and an unselected receiver uses zero. These give (6). Necessity is the substantive additional statement: applying Lemma 1 to (3) bounds the gain by $\sum_{j=1}^{r} s_j/[(1 - \beta/2)s_j + \nu]$, where s_j are the eigenvalues of $BB^\top$, including zeros. Strict concavity and monotonicity force $s_j = P/r$; the lemma then forces (7). Orthogonal changes of latent coordinates preserve the attainable risk because the channel noise is isotropic.

The fixed symmetry class. For graph consequences take either the cycle C_n, $4 \mid n$, with $\pi(v) = v + n/2 \pmod n$, or the Boolean cube $\{0,1\}^\ell$, $\ell \geq 2$, with $\pi(v) = v \oplus \mathbf{1}$. The group is cycle rotations or cube XOR translations, respectively. With $T_g e_v = e_{gv}$, require

$$T_g^\top G T_g = G \quad \text{for every translation } g. \tag{9}$$

This is equivalent to an orthogonal latent representation satisfying $BT_g = R_g B$. Write D_{inv} for the optimum under this constraint, still allowing arbitrary measurable decoders. We impose symmetry on the fixed encoder Gram, not on the internal operations of a decoder or on every possible graph architecture.

Theorem 3 (Exact fixed-invariant value)**.** *For the stated graphs and translation groups, under* (1) *with* $P > 0$ *and* $1 \le r \le n/2$*,*

$$D_{\mathrm{inv}} = n - \frac{rP}{(1-\beta r/n)P + r\nu}. \tag{10}$$

Consequently $D_{\mathrm{free}} < D_{\mathrm{inv}}$ *if* $r < n/2$*.*

Indeed invariance makes $Q_{vv} = q = \operatorname{tr} Q/n$. Positive semidefiniteness gives $|Q_{v,\pi(v)}| \le q$, and hence gain at most $T/(1-\beta T/n)$, where $T = \operatorname{tr} Q \le rP/(P+r\nu)$. An equal-energy projector onto any suitable real invariant rank-r subspace of the $+1$ eigenspace of J attains the bound. Complete Fourier sine/cosine pairs, with the constant and Nyquist modes, realize every stated cycle rank; even-weight Walsh characters do so on the cube. Appendix A supplies the exact rank counts and also characterizes invariant optimizers in either antipodal parity. Cube coordinate permutations are deliberately excluded: adding them can prevent these ranks from being attainable.

The same optimal value holds if the encoder may instead choose an arbitrary invertible latent representation satisfying $BT_g = R_g B$: Appendix A.5 symmetrizes the Gram within its invariant row space, preserving rank and power. This extends the value, not the optimizer classification or equivariance of the noisy channel.

The gap is an exact cost of requiring a *fixed* invariant Gram under local side information. With no local observation, both low-rank optima become the classical value $n - rP/(P + r\nu)$ (Schizas et al., 2007, Theorem 4), and every flat-spectrum rank-r projector is unrestricted-optimal; pair localization is no longer necessary. At $r = n/2$, the two values also agree, although their optimizer sets differ. Neither equality implies equal errors for fitted decoders. For $r = 0$ or $P = 0$, the zero-depth risk is n. With *exact* local observations, an additional genie argument gives, for all $0 \le r \le n$,

$$D_{\mathrm{free}} = n - \frac{2sP}{P + 2s\nu}, \qquad s = \min(r, n/2), \tag{11}$$

with zero gain when $sP = 0$. This high-rank extension is not asserted for noisy locality, and the low-rank all-Gram classification does not extend to $r > n/2$.

Theorem 4 (A sufficient positive-depth separation)**.** *Retain the Gaussian model and a transitive translation group. Receiver* v *now observes* $(x_w + \zeta_w)_{w \in A_v}$*, where* $v \in A_v$*,* $|A_v| = b$ *is independent of* v*, and* $\pi(v) \notin A_v$*. Suppose* $1 \le r \le n/2$*,* $P > 0$*, and set* $T_0 = rP/(P + r\nu)$*. If* $1 - \beta b T_0/n > 0$*, then*

$$D_{\mathrm{inv},A} \ge \max\left\{0,\ n - \frac{rP}{(1-\beta br/n)P + r\nu}\right\}. \tag{12}$$

The zero-depth pair decoder remains feasible and strictly beats every fixed invariant-Gram encoder whenever $br/n < 1/2$*.*

This is a sufficient separation certificate, not an exact positive-depth optimum. The proof bounds a local conditional-covariance correction by $bq^2/(d - bq)$; the stated denominator condition makes that step valid uniformly over feasible encoders. For a cycle k-hop ball, $b = 2k + 1$ while the antipode remains outside; thus matching the pair construction requires $k \ge \lceil n/(4r) - 1/2 \rceil$. On the cube, $b = \sum_{j=0}^{k} \binom{\ell}{j}$ with $k < \ell$. The result isolates neighborhood volume, not a universal depth lower bound for nonlinear GNNs. Supplementary consequences establish a spectral-projector task, independently randomized pair selection, and a target-only perturbation certificate under their stated assumptions.

4 From information floors to context-based decoders

A context-based predictor can use labeled examples, internal randomness, and interactions among query features. We therefore specify its entire information interface, rather than treating it as a rowwise linear regressor. Let M contain the encoder, graph, receiver identities, and declared canonicalization maps, together with any earlier training or independently sampled pair selection. Conditional on M, every context, validation, and fresh query row uses an independent *complete graph draw* from (1). Metadata exclude fresh targets, source/noise realizations, and seeds that reconstruct those realizations. Let C be permitted context and validation information, W independent decoder randomness, and

$$\mathcal{F} = \sigma(M, C, W, O_1, \dots, O_N). \tag{13}$$

Feature-only preprocessing or state from earlier feature-only calls is allowed when measurable with respect to this same information.

Proposition 5 (Independent-context and transductive transfer). *Write $\mu_i = \mathbb{E}[Y_i \mid M, O_i]$ and $\tau_i = \mathrm{Var}(Y_i \mid M, O_i)$ for the Gaussian posterior mean and variance. Every finite-risk, $\mathcal{F}$-measurable prediction $\widehat{Y}_i$ satisfies*

$$\mathbb{E}[(\widehat{Y}_i - Y_i)^2 \mid \mathcal{F}] = \tau_i + (\widehat{Y}_i - \mu_i)^2. \tag{14}$$

Thus expected risk equals the expected information floor plus expected decoder excess. It attains the floor exactly when $\widehat{Y}_i = \mu_i$ almost surely. Nonnegative receiver-weighted sums obey the same identity.

The Gaussian residual $Y_i - \mu_i$ is independent of its own observation; independent complete draws make it conditionally independent of the other permitted data. This proves the proposition even for predictions that depend jointly on all query features. It would fail if another query exposed a different vertex of the same graph draw: that vertex might directly reveal the target. The observed finite-batch MSE additionally contains oracle-residual sampling variation and a residual cross term; it is not identically floor plus sampled excess. Appendix B gives both identities and their statistical interpretation.

A defined canonicalized decoder. For an invariant encoder choose a reference receiver v_0 and a translation $g_v v_0 = v$. The wrapper supplies the fixed context/query algorithm with

$$O_v^{\mathrm{can}} = (u_v, R_{g_v}^\top z), \qquad BT_g = R_g B. \tag{15}$$

On a cycle take $g_v = v - v_0$; on the cube take $g_v = v \oplus v_0$. For equal-energy orthonormal rows, $R_g = (r/P)BT_gB^\top$. Apply the declared map to *each* context and query row, or use an independent reference-distribution context. The joint canonical feature/target law then equals that at v_0, without requiring the algorithm itself to be rotation invariant.

For the positive pair encoder, order selected pairs by a declared channel map. A selected receiver's wrapper moves its pair's channel to the first coordinate; unselected receivers form a separate stratum. Apply this permutation consistently to contexts and queries. Endpoint swaps and pair permutations preserve the white Gaussian law at zero depth. Hence the representative expected risks determine the mean receiver risk as

$$\frac{2r}{n} R_{\mathrm{selected}} + \left(1 - \frac{2r}{n}\right) R_{\mathrm{unselected}}. \tag{16}$$

At $r = n/2$, the unselected stratum is absent. These are risks of a defined wrapper; arbitrary unchanged PFN predictions across channel positions need not agree. Representative evaluation concerns mean risk, not the joint covariance of all receivers sharing one broadcast.

Finally, apply $\phi(t) = t + t^3/4$ coordinatewise *after* noise and canonicalization. This known Borel bijection preserves the observation sigma-field and therefore the exact Bayes floor, while changing the function a decoder must learn. The ordering is noise $\to$ canonicalization $\to$ transform. The real-arithmetic statement does not assert exact invertibility after float32 rounding; the inverse-transform control checks the float64 evaluation pipeline.

5 Separating representation loss from decoder error

The experiments measure how closely frozen decoders approach the representation-specific Bayes risks. A matched invertible presentation change isolates extraction difficulty, while dense encoder training tests recovery of the predicted pair structure. The frozen PFNs test extraction from these tabular context/query views without task-specific weight training; ridge provides the appropriate linear control for the raw Gaussian problem.

Design and inference interface. We evaluate TabPFN-3 and TabICLv2 regression checkpoints (Grinsztajn et al., 2026; Qu et al., 2026), alongside standardized ridge. The principal problem is C_{32} with antipodal targets, $r = P = 8$, broadcast-noise variance $\nu = 1/16$, and exact local observations. Each row uses a fresh complete Gaussian source and independent noise. Models receive context features and labels, then query features; query labels, conditional means, complete source vectors,

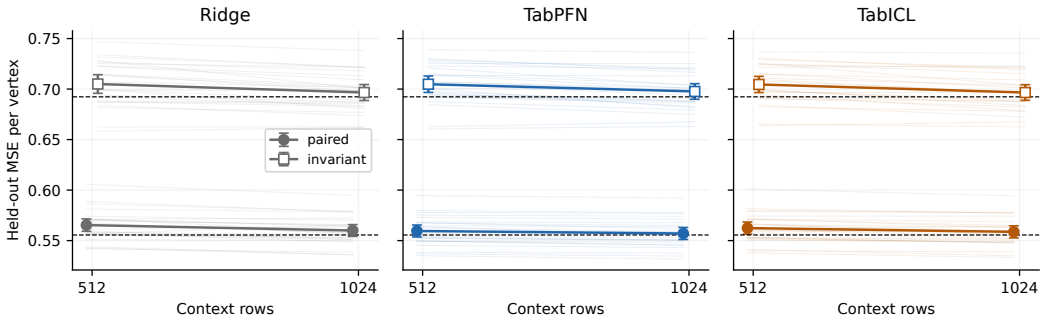


Figure 1: **One decoder, two information floors.** Held-out MSE on C_{32} in the 25 new replicates; bars are nominal 95% t intervals. Dashed lines are the exact per-vertex optima $5/9$ and $9/13$. Ridge is competitive on this linear-Gaussian task.

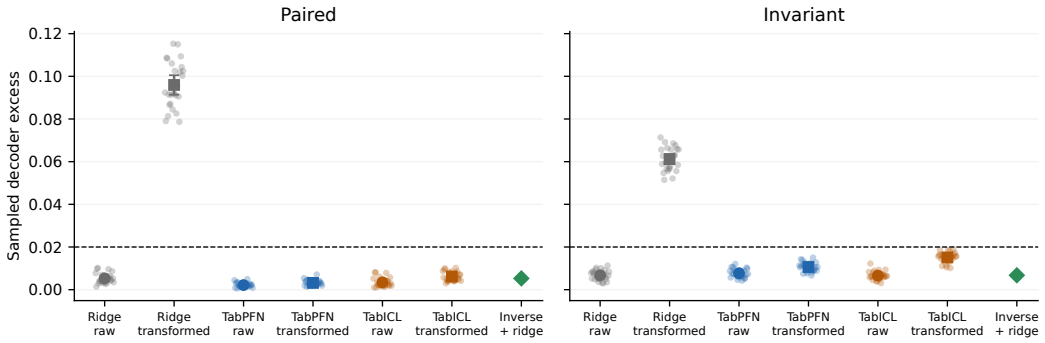


Figure 2: **Information preserved, extraction changed.** Sampled decoder excess at 1,024 context rows in the 25 new cycle replicates; bars are nominal 95% t intervals. The bijection $\phi(t) = t + t^3/4$ follows noise and canonicalization. The dashed 0.02 level is a frozen descriptive diagnostic. Explicit inversion restores raw-ridge predictions to numerical precision.

and seeds remain in the evaluation layer. Input-independent metadata canonicalize invariant channels or move a selected pair's active channel first, consistently across context and queries. Representative receivers have weights $2r/32$ and $1-2r/32$ for selected and unselected pairs; PFN rotation invariance is not assumed.

The primary replication cohort comprises 25 new independent context/query replicates, prospectively specified after the original five-replicate study. Each has 2,048 query rows, nested raw contexts of 512 and 1,024 rows, and 1,024 for other blocks. Checkpoints and inference settings match the original: one ensemble member, float32, six CPU threads, and 128-row query chunks. We retain the original five separately and evaluate ten cube replicates as a distinct secondary study. Query rows, strata, and technical repeats are not extra replicates. The two primary raw-1,024 PFN MSE contrasts use Bonferroni familywise 95% intervals (each 97.5%, 24 degrees of freedom); other intervals are nominal 95% paired or aggregate Student-t intervals.

Approaching two different floors. Sampled decoder excess is $N^{-1}\sum_i(\widehat{y}_i - \mu_i)^2$, where μ_i is the exact conditional mean. Expected risk equals the information floor plus expected excess; realized MSE also includes oracle-residual fluctuation and a cross term, so it can fall below a population floor. At 1,024 context rows both PFNs have small excess (Table 1). Their invariant-minus-paired MSE gaps are $0.140587\ [0.129777, 0.151398]$ for TabPFN and $0.137919\ [0.127279, 0.148559]$ for TabICL (familywise intervals), compared with the exact gap $16/117 = 0.136752$. Raw conditional means are linear; nonlinear decoding is unnecessary here. At 512 contexts, $7/100$ new-cohort PFN replicate aggregates exceed 0.02 (maximum 0.030512), despite all four group means and nominal upper limits

remaining below it. The original cohort has $2/20$ exceedances; separate group summaries appear in Appendix D.4.

Decoder	Raw paired	Raw invariant	Transformed paired	Transformed invariant
Ridge	0.00529	0.00675	0.09595	0.06120
TabPFN	0.00216	0.00770	0.00317	0.01062
TabICL	0.00346	0.00660	0.00616	0.01511

Table 1: Mean sampled decoder excess in the 25 new cycle replicates at 1,024 context rows. Floors are $5/9$ (paired) and $9/13$ (invariant), unchanged by the ideal real-valued transform. Full nominal intervals appear in the appendix.

Changing extraction difficulty. Coordinatewise $\phi(t) = t + t^3/4$ preserves the real-valued observation sigma-field and Bayes floor, but makes the displayed conditional mean nonlinear. Float32 conversion is a further deterministic processing step; exact bit-level invertibility is not assumed. Transformed ridge has much greater excess than either PFN. Inverse-transform-plus-ridge recovers raw predictions within 6.19×10^{-14} across the new cycle and cube float64 controls. All 16 cycle PFN model/arm/block groups at context $1{,}024$ have mean excess and nominal upper limits below the unchanged 0.02 diagnostic; all 400 constituent replicate aggregates also fall below it. This descriptive statement is neither a simultaneous attainment guarantee nor a general model ranking.

A lower Bayes floor need not imply smaller excess for an affine decoder. Both encoders satisfy $BB^\top = (P/r)I$, giving equal marginal broadcast covariance $(P/r + \nu)I$. The transformed-ridge reversal therefore depends on the joint feature–target structure, rather than greater paired-channel variance. In the declared native channel bases, a population calculation yields affine excess 0.088141 (paired) and 0.051479 (invariant), reproducing the direction before finite-context fitting (Appendix D.6); these are ideal function-class optima, not realized ridge errors.

Equal floors do not force equal fitted errors. No-local and rank-endpoint controls equalize both floors at $13/17$ and $1/5$, respectively; the endpoint uses $r = 16$ with unchanged $P = 8$. New no-local MSE gaps are 0.000653 $[-0.008692, 0.009997]$ for TabPFN and -0.001917 $[-0.011028, 0.007194]$ for TabICL. The original positive TabPFN endpoint gap, 0.014199 $[0.005734, 0.022665]$, is not reproduced: the new gap is 0.000461 $[-0.003016, 0.003938]$. TabICL's new endpoint gap is negative, -0.004145 $[-0.007685, -0.000605]$. These secondary differences include a shared oracle-batch difference -0.003554 $[-0.006982, -0.000127]$; TabPFN's excess and cross-term differences are 0.004492 and -0.000477. We preserve the cohort disagreement and paired component uncertainty (Appendix D; Figure 3). An interval containing zero does not establish equivalence.

A separate cube implementation check. Ten Q_5 replicates use XOR-antipodal pairing, the same principal budgets, and a frozen Walsh basis at context $1{,}024$, with raw and transformed inputs. All eight PFN model/arm/presentation group means and nominal upper excess limits are below 0.02, but one transformed-invariant TabICL replicate reaches 0.021086. The canonical raw white-Gaussian receiver laws agree up to latent orthogonal changes; channel orientation and the nonlinear presentation may affect practical decoding. These results check the specified cube construction, without isolating a topology effect or establishing real-graph robustness.

Learning and audit. The unchanged six dense population-objective fits attain their exact optima within 7×10^{-14}. Three validation-selected sample-trained fits have mean fixed-encoder Bayes risk 0.556075 and held-out MSE 0.557608, versus optimum 0.555556, with 0.168–0.176% off-pair Gram energy (Figure 4). These finite starts do not prove global convergence. The numerical audit reconstructs all $2{,}260$ new receiver records; all $1{,}230$ scientific receiver jobs completed. The charged pre-export workflow used 5.03 hours, with 1.232 GiB peak sampled process-tree-plus-supervisor RSS. Original validation anomalies remain historically unresolved despite byte-identical main-query replays. In the extension, same-chunk TabPFN validation predictions differ in $92/410$ fits (maximum 0.059510); all 410 TabICL same-chunk comparisons are exact. Changed-chunk probes also differ. Bounded fresh-process and settings diagnostics do not establish a general invariance guarantee or isolate a cause (Appendix D.10).

6 DISCUSSION AND LIMITATIONS

For $r < n/2$ with informative local observations, optimal fixed encoders concentrate on selected pairs despite the task's translations. Requiring encoder equivariance with an orthogonal latent action raises minimum risk, even over all measurable decoders. Averaging Grams preserves energy but can exceed the channel budget.

The scope is narrow by design. These conclusions assume white Gaussian inputs, Gaussian noise, and paired demands; correlated-source and heat-target controls reverse the ordering of the same two fixed codes, so pair concentration reflects this source and target rather than translation symmetry alone. The positive-depth result is a sufficient separation certificate, not an exact positive-depth optimum. Allowing an encoder-chosen invertible latent action gives the same optimal value (Proposition 6), without classifying every optimizer in that larger class. Nonorthogonal actions need not preserve isotropic noise, so this extension concerns deterministic encoder equivariance rather than equivariance of the noisy channel. Shared selection metadata and delivery costs are uncharged, and general nonlinear or long-block encoders enlarge the feasible set. Raw conditional means are linear, so the experiments imply neither a new PFN method nor a general advantage over ridge. The original five and new 25 cycle replicates differ on an endpoint control; the separate ten-replicate cube study does not isolate topology from channel orientation. Repeatability diagnostics cover one CPU environment and do not establish a cause for every anomaly. Finite encoder-training runs give no global convergence guarantee. The contribution is the exact optimization within the stated paired representation family; it is not a general priority claim for coding or symmetry breaking.

AI USE STATEMENT

OpenAI ChatGPT assisted with proof criticism, literature discovery and comparison, implementation, numerical auditing, visualization, and manuscript drafting. Separate AI-assisted proof and priority reviews informed the revisions. In a later editorial pass, Anthropic Claude revised prose in the abstract, introduction, related work, experimental narrative, and discussion, checked terminology consistency, and mapped the stated claims to their supporting theorems, tables, and cited sources. A separate Claude review checked proofs and numerical consistency. The editorial pass ran no experiments and altered no theorem, proof, numerical result, figure, or attribution. The pretrained-model experiments were executed locally by an author; their saved predictions were subsequently checked by numerical audit scripts. For this revision, private feedback from Google's Paper Assistant Tool (PAT), delivered through OpenReview, was used as an automated critique. ChatGPT assisted with feedback adjudication, manuscript edits, experiment extension code, analysis of the returned execution records, and reproducibility checks. In the final revision, ChatGPT also assisted with adjudicating the supplied reviews against the retained arrays and source, repairing and testing the offline audit, descriptive finite-batch analyses, and proving and checking the optimal-value extension to encoder-chosen invertible latent actions. These checks used the retained numerical records; no new pretrained-model inference was performed in that pass. The simulator and analysis implementation were AI-assisted; the archived observations and predictions arise from the documented seeded numerical computations and model executions, not text-generated experimental measurements. The mathematical arguments and numerical evidence are included for inspection. AI review is not independent human validation, and the authors are responsible for the submitted claims and artifacts.

REPRODUCIBILITY STATEMENT

Appendices A and B contain complete proofs and the decoder information contract. Appendix D specifies data generation, fixed model identities, inference settings, independent replicate analysis, and resource measurements. The accompanying anonymous supplement contains the frozen computational source, portable protocol information, saved scientific outputs, and offline audit scripts. Its manifest distinguishes unchanged numerical arrays from metadata sanitized for anonymous review. Pretrained weights are excluded; their identities and retrieval instructions are recorded. The 25 new cycle replicates are the primary replication cohort; the original five and the ten cube replicates are reported separately. Technical replays and repeatability diagnostics add no independent scientific replicates.

REFERENCES

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020. URL https://jmlr.org/papers/v21/19-322.html.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969, 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Pål Anders Floor, Anna N. Kim, Tor A. Ramstad, and Ilangko Balasingham. Zero delay joint source channel coding for multivariate Gaussian sources over orthogonal Gaussian channels. *Entropy*, 15(6):2129–2161, 2013. doi: 10.3390/e15062129. URL https://doi.org/10.3390/e15062129.

Jean Gallier. The Schur complement and symmetric positive semidefinite (and definite) matrices. University of Pennsylvania technical note, 2019. URL https://www.cis.upenn.edu/~jean/schur-comp.pdf. Version dated August 24, 2019.

Naveen Goela and Michael Gastpar. Reduced-dimension linear transform coding of correlated signals in networks. *IEEE Transactions on Signal Processing*, 60(6):3174–3187, 2012. doi: 10.1109/TSP.2012.2188716. URL https://arxiv.org/abs/1202.6299.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger,

Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report, 2026. URL https://arxiv.org/abs/2605.13986v2. Version 2.

Frank Hansen and Gert K. Pedersen. Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35:553–564, 2003. doi: 10.1112/S0024609303002200. URL https://arxiv.org/abs/math/0204049.

Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2308.10436.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15546–15566, 2023. URL https://proceedings.mlr.press/v202/kaba23a.html.

Sachin Katti, Shyamnath Gollakota, and Dina Katabi. Embracing wireless interference: Analog network coding. In *Proceedings of ACM SIGCOMM*, pp. 397–408, 2007. doi: 10.1145/1282380.1282425. URL https://groups.csail.mit.edu/netmit/6.888/www/papers/anc.pdf.

Hannah Lawrence, Vasco Portilheiro, Yan Zhang, and Sékou-Oumar Kaba. Improving equivariant networks with probabilistic symmetry breaking. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.21985v1.

Pablo Piantanida, Mari Kobayashi, and Giuseppe Caire. Analog index coding over block-fading MISO broadcast channels with feedback. In *2013 IEEE Information Theory Workshop*, 2013. doi: 10.1109/ITW.2013.6691245. URL https://doi.org/10.1109/ITW.2013.6691245.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.03336.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model, 2026. URL https://arxiv.org/abs/2602.11139v2. Version 2; published at ICML 2026.

Ioannis D. Schizas, Georgios B. Giannakis, and Zhi-Quan Luo. Distributed estimation using reduced-dimensionality sensor observations. *IEEE Transactions on Signal Processing*, 55(8):4284–4299, 2007. doi: 10.1109/TSP.2007.895987. URL https://doi.org/10.1109/TSP.2007.895987.

Shuo Tan and Syed A. Jafar. Wireless linear computation broadcast, 2026. URL https://arxiv.org/abs/2608.05692v1. Version 1.

Roy Timo, Alex Grant, and Gerhard Kramer. Lossy broadcasting in two-way relay networks with common reconstructions, 2010. URL https://arxiv.org/abs/1011.4725v1. Version 1; ordinary complementary delivery is Example 2.

Fanggang Wang, Xiaojun Yuan, Soung Chang Liew, and Dongning Guo. Wireless MIMO switching: Weighted sum mean square error and sum rate optimization. *IEEE Transactions on Information Theory*, 59(9):5297–5312, 2013. doi: 10.1109/TIT.2013.2259893. URL https://arxiv.org/abs/1207.6706v2.

A Complete proofs and boundary cases

A.1 Posterior conditioning among all measurable decoders

Set $C = BB^\top + \nu I_r$ and $G = B^\top B$. For nonzero h, $h^\top C h = \|B^\top h\|^2 + \nu\|h\|^2 > 0$. A singular value decomposition proves $B^\top C^{-1} B = G(G + \nu I_n)^{-1}$, with eigenvalues $s/(s+\nu) \in [0, 1)$; thus (2) holds even if B has deficient rank. Define $m_z = B^\top C^{-1} z$ and $e = x - m_z$. Direct expansion gives

$$\mathrm{Cov}(e, z) = 0, \qquad \mathrm{Cov}(e) = I_n - Q.$$

The pair (e, z) is jointly Gaussian. Its joint characteristic function factors because the covariance is block diagonal, so e and z are independent. Consequently $m_z = \mathbb{E}[x \mid z]$ and $\mathrm{Cov}(x \mid z) = I_n - Q$.

Fix $j = \pi(v) \neq v$. After observing z, the local innovation is $u_v - (m_z)_v = e_v + \zeta_v$, with variance $d - Q_{vv} > 0$ and covariance $-Q_{jv}$ with e_j. A second Gaussian residual calculation gives

$$\mu_v = (m_z)_j - \frac{Q_{jv}}{d - Q_{vv}}\big(u_v - (m_z)_v\big), \tag{17}$$

$$\tau_v = 1 - Q_{jj} - \frac{Q_{jv}^2}{d - Q_{vv}}. \tag{18}$$

The denominator is positive because $d \geq 1$ and $Q_{vv} < 1$. The $\{j, v\}$ principal block of $I_n - Q$ is positive definite; its Schur complement proves $\tau_v > 0$ at $\sigma^2 = 0$, and increasing the local variance only reduces the subtraction. Every finite-risk measurable $f_v(O_v)$ obeys

$$\mathbb{E}[(f_v - Y_v)^2] = \tau_v + \mathbb{E}[(f_v - \mu_v)^2],$$

since the conditional expectation of $Y_v - \mu_v$ given O_v is zero. Equality holds exactly when $f_v = \mu_v$ almost surely. Infinite-risk decoders cannot improve the infimum. Summing over the permutation π proves (3).

Why noise variances alone do not suffice. Let $n = 2, r = 1$, $B = (a, a)$, $a > 0$, and $\sigma^2 = 0$, but replace the Gaussian channel noise by an independent uniform sign times $\sqrt{\nu}$. After subtracting its exact local value, each receiver observes $W = aX + \eta$ for a standard normal target X. Bayes' formula gives

$$\mathbb{E}[X \mid W = w] = \frac{w - \sqrt{\nu}\tanh(\sqrt{\nu}\, w/a^2)}{a}.$$

This differs from the best linear predictor $aw/(a^2 + \nu)$ on a set of positive probability. Orthogonality makes its MMSE strictly smaller than $\nu/(a^2 + \nu)$, the Gaussian scalar-channel value. The source and both noise laws in (1) therefore belong in every self-contained statement of the theorem.

A.2 Proof of Lemma 1

Reorder coordinates into their fixed pairs. A principal pair block has the form $A = \left(\begin{smallmatrix} u & w \\ w & v \end{smallmatrix}\right)$, with eigenvalues $a, b \in [0, 1)$. Put

$$u = (a+b)/2 + t, \quad v = (a+b)/2 - t, \quad c = (a-b)^2/4, \quad w^2 = c - t^2, \quad h = d - (a+b)/2.$$

Then $0 \leq t^2 \leq c$, $h > 0$, and $h^2 - c = (d-a)(d-b) > 0$. Define

$$L(A) = w^2\left(\frac{1}{d-u} + \frac{1}{d-v}\right) = \frac{2h(c - t^2)}{h^2 - t^2}, \qquad f(s) = \frac{s^2}{2d - s}.$$

The two exact scalar remainders are

$$\frac{2c}{h} - L(A) = \frac{2t^2(h^2 - c)}{h(h^2 - t^2)} \geq 0, \tag{19}$$

$$f(a) + f(b) - \frac{2c}{h} = \frac{2ab\{2(d-a)(d-b) + d(2d-a-b)\}}{(2d-a)(2d-b)(2d-a-b)} \geq 0. \tag{20}$$

All displayed denominator factors and the bracket in the second numerator are positive. The first remainder vanishes exactly when $t = 0$, and the second exactly when $ab = 0$. These statements

include $a = b = 0$ and remain valid at $d = 1$. Thus $L(A) \le \operatorname{tr} f(A)$, with equality exactly for zero blocks or the balanced rank-one blocks in (5).

It remains to pass from principal pair blocks to the full matrix. This inverse-compression step is standard Schur-complement algebra (Gallier, 2019); operator Jensen inequalities provide a more general setting for such compressions (Hansen & Pedersen, 2003). We give the needed equality argument directly. Let $M = 2dI_n - Q \succ 0$ and, for one pair with nonempty complement, write

$$M = \begin{pmatrix} C & E \\ E^\top & D \end{pmatrix}, \qquad C = 2dI_2 - A, \qquad S = C - ED^{-1}E^\top \succ 0.$$

Block elimination gives $(M^{-1})_{jj} = S^{-1}$. Since $0 \prec S \preceq C$, conjugating by $C^{-1/2}$ and inverting shows $S^{-1} \succeq C^{-1}$. Using $f(s) = 4d^2/(2d - s) - 2d - s$, the linear and constant parts cancel in the sum over all pair blocks A_j:

$$\operatorname{tr} f(Q) - \sum_j \operatorname{tr} f(A_j) = 4d^2 \sum_j \operatorname{tr}\big[(M^{-1})_{jj} - M_{jj}^{-1}\big] \ge 0. \tag{21}$$

Every bracket is positive semidefinite. If the sum has zero trace, every bracket is the zero matrix. Hence $S^{-1} = C^{-1}$ and $S = C$, so $ED^{-1}E^\top = 0$. The latter is $(ED^{-1/2})(ED^{-1/2})^\top$, forcing $E = 0$. Applying this to every pair removes all cross-pair couplings. For $n = 2$ the complement is empty and the compression step is vacuous. Zero pair blocks introduce no exception: $|Q_{ij}|^2 \le Q_{ii}Q_{jj}$ implies that a zero diagonal entry of a positive semidefinite matrix has an entirely zero row and column. Summing the three nonnegative remainders proves the inequality and every asserted equality condition. Conversely, the prescribed block diagonal family makes every remainder zero. □

A.3 Proof of Theorem 2

Let $s_1, \dots, s_r \ge 0$ be the eigenvalues of $BB^\top$, including zeros, and $q_j = s_j/(s_j + \nu)$. They are squared singular values, not singular values. From (3) and the paired-entry lemma,

$$n - D(B) \le \sum_{j=1}^{r} \left(q_j + \frac{q_j^2}{2d - q_j} \right) = \sum_{j=1}^{r} \frac{s_j}{(1 - \beta/2)s_j + \nu}. \tag{22}$$

For $c = 1 - \beta/2 \in [1/2, 1)$, $h(s) = s/(cs + \nu)$ has

$$h'(s) = \frac{\nu}{(cs + \nu)^2} > 0, \qquad h''(s) = -\frac{2c\nu}{(cs + \nu)^3} < 0.$$

With $p = \sum_j s_j \le P$, Jensen's inequality and monotonicity give

$$\sum_j h(s_j) \le rh(p/r) \le rh(P/r) = \frac{rP}{(1 - \beta/2)P + r\nu}.$$

The positive pair code and decoder (8) attain this gain. A selected receiver's explained variance is $a^2/[(2 - \beta)a^2 + \nu]$, there are $2r$ selected receivers, and the unselected targets are independent of their observations.

An optimizer must attain every inequality. Strict monotonicity forces $p = P$; strict concavity forces all $s_j = P/r > 0$ (for $r = 1$, the latter statement follows directly). The nonzero eigenvalues of Q are therefore all $q_* = P/(P + r\nu)$, and its rank is r. Lemma 1 forces exactly r nonzero balanced rank-one pair blocks. The inverse spectral map

$$G = \nu Q(I_n - Q)^{-1}$$

preserves eigenspaces and maps q_* to P/r, yielding exactly (7). Conversely any such Gram attains every step. All full-row-rank real factors of the same Gram differ by a left orthogonal transformation, by their singular value decompositions. For a signed pair row with coefficients $b_v, b_j \in \{-a, a\}$, the attaining decoder is

$$\widehat{x}_j = \frac{b_j}{(2 - \beta)a^2 + \nu}(z_{\text{pair}} - \beta b_v u_v).$$

Thus both signs, not just the displayed positive construction, attain the value. □

A.4 Invariant Grams and their exact optimum

Latent representation, including its group law. Suppose $T_g^\top G T_g = G$. If $Bx = 0$, then $\|BT_g x\|^2 = x^\top T_g^\top G T_g x = x^\top G x = 0$. Consequently the linear map $R_g(Bx) = BT_g x$ is well-defined on $\operatorname{im} B$. It preserves inner products by polarization and has inverse $R_{g^{-1}}$. Further, $R_g R_h(Bx) = BT_g T_h x = BT_{gh}x = R_{gh}(Bx)$. Extending every R_g by the identity on the common orthogonal complement gives an orthogonal representation on $\mathbb{R}^r$. The converse follows by taking Grams of $BT_g = R_g B$.

Value. Gram invariance passes to $Q = G(G + \nu I)^{-1}$. Transitivity makes its diagonal constant, $Q_{vv} = q = T/n$ for $T = \operatorname{tr} Q$. A positive semidefinite 2×2 minor implies $|Q_{v,\pi(v)}| \le q$, giving

$$n - D(B) \le nq + \frac{nq^2}{d - q} = \frac{T}{1 - \beta T/n}.$$

Here $d - q > 0$ since $q < 1$. The right side increases strictly with T on the feasible interval. Monotonicity and strict concavity of $s/(s + \nu)$ imply

$$T = \sum_j \frac{s_j}{s_j + \nu} \le \frac{rP}{P + r\nu} = T_0.$$

This yields the lower bound in (10); its denominator is positive since $\beta r/n \le 1/2$. To attain it take a real translation-invariant orthogonal projector Π of rank r in the $+1$ eigenspace of J and set $G = (P/r)\Pi$. Then $Q = \alpha\Pi$, $\alpha = P/(P + r\nu)$, has diagonal $\alpha r/n$ and $QJ = Q$, so every paired entry equals that diagonal. Any orthonormal row basis of $\operatorname{im}\Pi$, scaled by $\sqrt{P/r}$, is a feasible factor B. Thus all inequalities are equalities.

Every required real rank. On C_n with $n = 4k$, a complex Fourier frequency j has antipodal multiplier $(-1)^j$. The real plus-parity subspace consists of the constant and Nyquist singleton modes $j = 0, n/2$, and $k - 1$ complete real sine/cosine planes. For even $r < n/2$, use $r/2 \le k - 1$ planes. For odd $r < n/2$, use one singleton and $(r - 1)/2 \le k - 1$ planes. For $r = n/2$, use the entire plus space. If $n = 4$, there are no planes; the two singleton modes give $r = 1$ and $r = 2$ directly. No conjugate pair is split. On the cube $\{0, 1\}^\ell$, the real Walsh character $\chi_a(v) = (-1)^{a\cdot v}/\sqrt{n}$ obeys $T_g\chi_a = (-1)^{a\cdot g}\chi_a$ and $J\chi_a = (-1)^{|a|}\chi_a$. There are $n/2$ even-weight characters, and any r of them give the required projector.

Complete invariant equality classification. Strictness in the trace bounds forces full power and $G = (P/r)\Pi$, with Π a translation-invariant rank-r projector; this step also applies in the separate no-local model. For $\beta > 0$, equality in the paired-entry bounds additionally requires $|Q_{v,\pi(v)}| = q > 0$ for every vertex. If $a_v = Q^{1/2}e_v$, these are equality cases of Cauchy–Schwarz with equal norms, so $a_{\pi(v)} = \epsilon_v a_v$. The paired entries are constant under translations because π commutes with the group and the action is transitive. Thus all ϵ_v have one common value $\epsilon \in \{-1, 1\}$. Equivalently $QJ = \epsilon Q$, so $\operatorname{im}\Pi$ lies entirely in one antipodal parity. Conversely every invariant rank-r projector in either parity, with scaling P/r, attains the value. On the stated cycles minus parity has only sine/cosine planes and therefore admits only even ranks. Cubes admit either parity at every rank $r \le n/2$.

The strict gap follows because the invariant gain denominator exceeds the free denominator by $\beta P(1/2 - r/n) > 0$ when $r < n/2$. At $r = n/2$, invariant optimizers are the scaled projectors $(I_n + J)/2$ and $(I_n - J)/2$, whereas unrestricted optimal Grams also allow different signs on different pairs. Equal values do not identify the optimizer sets.

Why the cube group must be specified. For the three-dimensional cube, imposing coordinate permutations in addition to XOR translations forces projector eigenvalues to be constant on the Walsh weight shells, whose dimensions are $1, 3, 3, 1$. An equality-case rank-two projector would need to lie in one antipodal parity. Neither parity has a union of shells of total dimension two. The exact lower bound is therefore unattainable in this enlarged group. The feasible Gram set is compact and the risk continuous, so its optimum is strictly worse. The XOR-only scope is substantive.

A.5 AN OPTIMAL-VALUE EXTENSION TO INVERTIBLE LATENT ACTIONS

Proposition 6 (Symmetrization on an invariant row space)**.** *Under* (1)*, let a finite group act by permutation matrices T_g and commute with J. With the same channel and power budgets, the optimum over encoders admitting some representation $R_g \in \mathrm{GL}(r, \mathbb{R})$ with $BT_g = R_g B$ equals the invariant-Gram optimum. The representation may be chosen with B. Transitivity is not required for this equality of values.*

Proof. Such an equivariant B has invariant kernel, hence invariant row space $V = (\ker B)^\perp$, because the T_g are orthogonal and include their inverses. Conversely, if V is invariant, the rule $R_g(Bx) = BT_g x$ is well defined on $\mathrm{im}\, B$, is invertible, and obeys the group law. Extend it by the identity on the fixed complement $(\mathrm{im}\, B)^\perp$. This proves the equivalence, including deficient rank.

For $\sigma^2 > 0$, Gaussian conditioning expresses the risk as

$$D_\sigma(G) = \sum_v e_{\pi(v)}^\top \left(I_n + G/\nu + e_v e_v^\top/\sigma^2\right)^{-1} e_{\pi(v)}.$$

For positive definite A, $a^\top A^{-1} a = \sup_w (2a^\top w - w^\top A w)$, a supremum of affine functions of A. Thus D_σ is convex in G. Reindexing receivers, using $T_g J = J T_g$ and the common local-noise variance, gives $D_\sigma(T_g G T_g^\top) = D_\sigma(G)$. For any feasible invariant-row-space encoder, put

$$\bar{G} = |\Gamma|^{-1} \sum_{g \in \Gamma} T_g G T_g^\top.$$

This Gram is positive semidefinite, invariant, and has the same trace as G. Every summand is supported on the same invariant space V, so $\mathrm{rank}\, \bar{G} \le \dim V \le r$. It therefore has a factor with r rows, and convexity gives $D_\sigma(\bar{G}) \le D_\sigma(G)$.

For exact locality, set $K = (I_n + G/\nu)^{-1} \succ 0$. The equivalent conditional-covariance expression is

$$D_\sigma(G) = \sum_v \left(K_{\pi(v),\pi(v)} - \frac{K_{\pi(v),v}^2}{\sigma^2 + K_{vv}} \right).$$

It is continuous at $\sigma^2 = 0$, since $K_{vv} > 0$. Taking that limit for the fixed pair $G, \bar{G}$ preserves the inequality; no limit is interchanged with optimization. Every encoder in the enlarged class therefore has an invariant-Gram competitor of no greater risk. The reverse inequality between infima follows from inclusion of the invariant-Gram class. For the separate no-local model, the same argument applies directly to the convex risk $\mathrm{tr}(I_n + G/\nu)^{-1}$. □

Consequently Theorem 3's value also holds for this enlarged class on the stated graphs and ranks. This does not classify all its optimizers or concern a prescribed nonorthogonal representation. Nor does it make the noisy channel equivariant: $R_g \eta$ has covariance $\nu R_g R_g^\top$, generally different from νI_r. Rank preservation in this proof relies on the starting row space being invariant; it does not apply to an arbitrary pair-localized optimum.

A.6 BOUNDARY REGIMES

No local observation. Here $D(B) = n - \mathrm{tr}\, Q$; the parameter $\beta = 0$ denotes a separate observation model, not a finite local-noise variance. For $1 \le r \le n$, the strict trace argument gives

$$D_{\mathrm{free}} = n - \frac{rP}{P + r\nu}, \qquad G = (P/r)\Pi$$

for *every* orthogonal rank-r projector Π, and these are all optimal Grams when $P > 0$. The local paired-entry term has disappeared, so no localization or antipodal-parity necessity remains. On the stated graph families invariant projectors realize these ranks: the cycle has two real singleton modes and complete two-dimensional planes, and the cube has real one-dimensional characters. Thus the invariant value equals the free value. At zero power or zero channel count the zero-depth target is independent of the permitted local observation, and the total risk is n; formulas with zero denominators are interpreted separately.

Exact locality at every channel count. For $r \le n/2$, equation (11) is Theorem 2 with $\beta = 1$. For $r > n/2$, give the receiver demanding x_j all other source coordinates. This can only lower its risk, and leaves the scalar Gaussian channel $b_j x_j + \eta$ after subtraction, where b_j is column j of B. Therefore

$$D(B) \ge \sum_{j=1}^{n} \frac{\nu}{\nu + \|b_j\|^2} \ge \frac{n^2\nu}{P + n\nu}.$$

The second inequality uses strict convexity and monotone decrease of $\nu/(\nu + t)$: equality in this step needs equal column energies and full power. Coding all $n/2$ pairs with coefficients $\sqrt{P/n}$ and zeroing unused latent rows attains the bound. This proves the value but does not classify all high-rank optimizers. For instance $B = \sqrt{P/n}\, I_n$ also attains it at $r = n$, yet has full-rank Gram. Noisy locality changes this question: at $n = 2$, $P = 2$, $\nu = 1$, $\sigma^2 = 1$, the one-channel optimum is $6/5$, whereas $B = I_2$ with two channels gives total risk 1. Thus saturation at $n/2$ cannot be extended to noisy local observations.

A.7 Proof of Theorem 4

For $w = \pi(v) \notin A_v$, condition first on z. The conditional target variance is $1 - q$, its covariance with the local observations is $-Q_{w,A_v}$, and their conditional covariance is $dI - Q_{A_v,A_v} \succ 0$. Hence receiver risk is exactly

$$1 - q - Q_{w,A_v}(dI - Q_{A_v,A_v})^{-1} Q_{A_v,w}.$$

The target-exclusion condition removes an otherwise present identity term from this cross covariance. Since $Q \succeq 0$ has constant diagonal,

$$0 \preceq Q_{A_v,A_v} \preceq bqI, \qquad \|Q_{w,A_v}\|^2 \le bq^2.$$

Under $d > bq$, the reduction is at most $bq^2/(d - bq)$. Summing and setting $T = nq$ bounds the total gain by $T/(1 - \beta bT/n)$. The hypothesis $1 - \beta bT_0/n > 0$ makes this estimate valid and increasing throughout $0 \le T \le T_0$. Substitution yields (12); nonnegativity of risk permits truncating the lower bound at zero. If $br/n < 1/2$, the denominator condition follows automatically and the invariant gain denominator strictly exceeds $(1 - \beta/2)P + r\nu$. Since $v \in A_v$, the zero-depth decoder remains feasible by ignoring extra local observations, proving the strict comparison. No exact positive-depth optimum follows. □

A.8 Additional consequences with explicit scope

Proposition 7 (An interleaved spectral-projector task)**.** *On the stated cycles and cubes, $H = (I_n + J)/2$ is an orthogonal graph spectral projector. With exact local x_v, the optimized risk of predicting Hx in either fixed encoder class is one quarter of the corresponding risk of predicting Jx.*

Proof. The identities $J = J^\top$ and $J^2 = I_n$ imply $H = H^\top = H^2$. On the cycle with combinatorial Laplacian L, the multiplier of $I_n - L/2$ at Fourier index j is $\cos(2\pi j/n)$. Thus the Chebyshev polynomial of degree $n/2$ gives $J = T_{n/2}(I_n - L/2)$, with multiplier $(-1)^j$. Conjugate frequencies have the same parity, so H selects entire real Laplacian eigenspaces. On the cube the Laplacian eigenvalue is $2|a|$ and H selects even-weight Walsh eigenspaces. The maps $f_v \mapsto (x_v + f_v)/2$ and $h_v \mapsto 2h_v - x_v$ are inverse bijections between admissible predictors for the two tasks. They scale errors by $1/2$ pointwise, hence squared errors by $1/4$, for every encoder before taking infima. Exact locality is needed. This interleaved band is not a usual low-pass band, and the pair construction itself lies in the selected band. □

Proposition 8 (Independent pair randomization)**.** *Choose r of the $n/2$ pairs uniformly and independently of fresh source and noise draws. Disclose the selected set and channel assignment to the receivers, and use the positive pair Bayes decoder. The total expected risk is D_{free}, every receiver has expected risk D_{free}/n, and the prediction law averaged over selection and noise is equivariant under the stated translations.*

Proof. Conditional on each selected set, the encoder attains Theorem 2. A receiver's pair is selected with probability $2r/n$, yielding the same marginal mixture of selected and unselected risks at every vertex. Translations permute the uniform selected set and its channel assignment. The induced

channel permutation preserves the isotropic noise law; the source and local-noise laws are invariant too. Changing variables in this finite mixture proves distributional equivariance, including as a prediction kernel conditional on a fixed input. Random selection does not improve the unrestricted optimum, and does not satisfy a single fixed-Gram constraint. Shared selection/channel metadata are not charged to the real channel budget. Their independence from the fresh source prevents them from carrying extra target information. □

Proposition 9 (Target-only perturbation certificate)**.** *Fix an admissible observation and predictor class $\mathcal{C}$, with a finite-risk predictor, and write $D_{\mathcal{C}}(H)$ for its optimal total squared error for Hx. For a white source,*

$$\left|\sqrt{D_{\mathcal{C}}(H)} - \sqrt{D_{\mathcal{C}}(H_0)}\right| \le \|H - H_0\|_F.$$

If two such fixed classes have a root-risk gap $g > 0$ at H_0, their ordering persists when $\|H - H_0\|_F < g/2$.

Proof. For any common admissible predictor $\widehat{Y}$, the reverse triangle inequality in L^2 bounds the change in its root error by $\|(H - H_0)x\|_{L^2} = \|H - H_0\|_F$. Choose arbitrarily near-optimal predictors and take infima in both directions; existence of an optimizer is unnecessary. Apply this bound to both classes to obtain new gap at least $g - 2\|H - H_0\|_F$. This comparison fixes the graph, source law, observations, and class; it does not justify changing those objects together with the target. □

B Decoder transfer, canonicalization, and finite-batch risk

B.1 Permitted data and the conditional residual argument

Here an independent draw means a complete triple $(x^{(i)}, \zeta^{(i)}, \eta^{(i)})$ with the laws in (1). A row is not a second receiver observing the same realized graph. The task metadata M may include B, the graph, pairing, receiver identities, channel assignments, canonicalization maps, independent pair-selection randomness, and earlier training history. When B was learned, condition on that earlier history: it must not use the fresh queries. Conditional on M, fresh query draws have the stated Gaussian laws, are mutually independent, and are independent of C, the permitted contexts, validation data, and learned state. Internal randomness W is independent of these fresh draws as well.

In particular, M, C, and W contain neither fresh labels nor any fresh source or noise values. A source/noise generation seed from which those values can be reconstructed is also excluded. In contrast, an independent seed used only to select pairs is allowed. An archive may store generation seeds for reproducibility without making them inference inputs. Pseudorandom streams instantiate the intended stochastic model; different hash-derived seeds alone are not a theorem of mathematical independence.

For fixed M, let

$$K_i = \mathrm{Cov}(O_i \mid M), \qquad c_i = \mathrm{Cov}(O_i, Y_i \mid M),$$
$$\mu_i = c_i^\top K_i^{-1} O_i, \qquad \tau_i = \mathrm{Var}(Y_i \mid M) - c_i^\top K_i^{-1} c_i.$$

The conditional means of O_i, Y_i are zero under the model. The observation covariance is positive definite: for scalar a and vector b, writing e_v for a coordinate basis vector,

$$\mathrm{Var}(a u_v + b^\top z \mid M) = \|a e_v + B^\top b\|^2 + a^2\sigma^2 + \nu\|b\|^2.$$

Vanishing variance forces $b = 0$ because $\nu > 0$, and then $a = 0$. With no local observation $K_i = BB^\top + \nu I_r \succ 0$ instead.

Let $\epsilon_i = Y_i - \mu_i$. It is jointly Gaussian with O_i and $\mathrm{Cov}(\epsilon_i, O_i \mid M) = 0$ by construction. The Gaussian characteristic function factors, proving conditional independence of ϵ_i and O_i, not just zero correlation. Independence of the complete draws makes the pair (ϵ_i, O_i) conditionally independent of (C, W, O_{-i}). Consequently, for $\mathcal{F}$ in (13),

$$\mathbb{E}[\epsilon_i \mid \mathcal{F}] = 0, \qquad \mathbb{E}[\epsilon_i^2 \mid \mathcal{F}] = \tau_i.$$

If M is random, its value can affect τ_i; unconditional independence of the residual from M is not needed or claimed. For $\mathcal{F}$-measurable $\widehat{Y}_i$, expansion of $(\widehat{Y}_i - \mu_i - \epsilon_i)^2$ proves Proposition 5. Taking expectations gives

$$\mathbb{E}[(\widehat{Y}_i - Y_i)^2] = \mathbb{E}[\tau_i] + \mathbb{E}[(\widehat{Y}_i - \mu_i)^2].$$

The excess is nonnegative and vanishes exactly when $\widehat{Y}_i = \mu_i$ almost surely. Conditional or unconditional nonnegative weighted sums preserve the identity when the weights are declared from metadata. Query-feature interactions and stateful feature-only calls remain covered if all predictions are measurable in the same $\mathcal{F}$.

A same-graph counterexample. Take $n = 2$, $B = 0$, and exact local observations. Receiver 1's target is x_2, and its own observation x_1 leaves risk 1. If the query table also contains receiver 2 from this same graph draw, that row exposes x_2 and reduces the first prediction risk to zero. Unique row IDs do not prevent this information leak. Independent complete graph draws are therefore part of the proposition, not an incidental implementation choice.

B.2 The exact finite-batch decomposition

Put $\delta_i = \widehat{Y}_i - \mu_i$ and $\epsilon_i = Y_i - \mu_i$. For every realized batch,

$$\begin{aligned}\widehat{\mathrm{MSE}} &= \frac{1}{N}\sum_i (\widehat{Y}_i - Y_i)^2 \\ &= \frac{1}{N}\sum_i \epsilon_i^2 + \frac{1}{N}\sum_i \delta_i^2 - \frac{2}{N}\sum_i \delta_i \epsilon_i. \end{aligned} \tag{23}$$

The first term is the oracle's realized error, the second is sampled decoder excess, and the last is a residual cross term. Conditioning on the allowed feature information yields

$$\widehat{R}_{\mathrm{cond}} := \frac{1}{N}\sum_i \big(\tau_i + \delta_i^2\big) = \mathbb{E}[\widehat{\mathrm{MSE}} \mid \mathcal{F}]. \tag{24}$$

Thus floor plus sampled excess is an unbiased estimator of the expected MSE under the protocol and is above the corresponding mean floor by construction. It is distinct from the observed MSE, which may lie below the population floor in a finite batch. When receiver floors vary, use the average τ_i rather than inserting an unrelated constant. The same algebra applies to declared selected/unselected receiver weights.

Shared contexts and transductive processing can correlate losses across query rows. Within-batch iid standard errors consequently need separate justification and are only diagnostics otherwise. Independent complete context/query-batch replicates are the appropriate units for the reported replicate summaries. A five-replicate Student interval is nominal, not an exact coverage guarantee under arbitrary replicate distributions. Replaying an identical dataset does not enlarge that sample size. Equal optimal floors at the no-local or rank endpoint controls also do not imply equal fitted-decoder risk: their excesses can differ, and realized oracle and cross terms can differ.

B.3 Invariant-receiver canonicalization

Use $T_g e_v = e_{gv}$ and the orthogonal representation already proved. Fix v_0 and choose g_v with $g_v v_0 = v$; use the identity for $v = v_0$. Set $x' = T_{g_v}^\top x$. Because x is white Gaussian, x' has the same law as x, while

$$x'_{v_0} = x_v, \qquad x'_{\pi(v_0)} = x_{\pi(v)}, \qquad R_{g_v}^\top z = Bx' + R_{g_v}^\top \eta.$$

The last identity uses $R_g^\top B = B T_g^\top$, and the middle one uses the fact that π commutes with translations. The rotated noise is independent isotropic Gaussian. Therefore

$$((u_v, R_{g_v}^\top z), Y_v) \overset{\text{law}}{=} ((u_{v_0}, z), Y_{v_0}). \tag{25}$$

The result holds jointly for independent context and query draws and for a common independent internal random seed. Define the decoder wrapper to canonicalize every row before applying the same algorithm to its context/query tables. Alternatively, fix an independent context already sampled at v_0

and canonicalize every query row. In either case, the entire input/target law for each representative experiment matches. For a mixed receiver query table, apply each row's declared map and keep its position fixed; the rows remain independent with the reference law. This proves equality of wrapper risks without assuming any invariance property of the underlying algorithm.

For $BB^\top = (P/r)I_r$, one may compute

$$R_g = (r/P)BT_gB^\top.$$

Indeed $\Pi = (r/P)B^\top B$ is the invariant row-space projector and commutes with T_g; substitution proves $R_gB = BT_g$, $R_gR_h = R_{gh}$, and $R_gR_g^\top = I_r$. The explicit vertex choices are $g_v = v - v_0$ modulo n on a cycle and $g_v = v \oplus v_0$ on a cube. Channel ordering is fixed in the row basis of B. At v_0 the action is the identity. A representative experiment therefore samples the wrapper's reference law without exercising every nontrivial vertex map; that is distinct from an all-vertex software verification.

Why an unchanged decoder need not share the symmetry. On C_4, index vertices by $0, 1, 2, 3$ and take $B = a(1, -1, 1, -1)$, $a = \sqrt{P}/2 > 0$. Its Gram is rotation invariant, since one rotation negates its row. But the unchanged decoder $\delta(u, z) = z$ has risk

$$\mathbb{E}[(z - x_{v+2})^2] = 1 + P + \nu - 2a(-1)^v,$$

which differs between even and odd vertices. Canonicalization multiplies z by $(-1)^v$ and removes this discrepancy. Invariant encoder Grams therefore do not automatically make arbitrary PFNs, or arbitrary measurable decoders, invariant.

B.4 Selected and unselected pair representatives

Consider the positive pair encoder with a declared ordering of its r selected pairs. A selected receiver identifies its pair's channel j; choose the permutation matrix P_j that exchanges coordinates j and 1 (the identity if $j = 1$), and supply $(u_v, P_j z)$ to the algorithm. Use the same rule for each context row and query row, with labels left as their paired targets. Within a selected pair, swapping endpoints preserves the joint Gaussian source/target law. Across selected pairs, permute entire pair blocks together with the assigned broadcast channels. This maps any selected receiver to the first-channel reference law. These source permutations are valid at zero depth because only the pairing and white source enter the statistical model; they need not be automorphisms of a graph with larger local neighborhoods.

An unselected pair is independent of all broadcast coordinates. All unselected receivers therefore share their own reference law, with target independent of the permitted local feature and broadcast. Keep selected/unselected status as declared task metadata and evaluate the two strata separately. Counting their vertices gives exactly (16). At the endpoint $r = n/2$, its second weight is zero and there is no unselected stratum. The reduction proves expected mean/sum risk; separately sampled representative rows do not recover simultaneous cross-receiver error covariance from a common broadcast draw. It also does not assert identical losses for an unchanged column-order-sensitive algorithm before the wrapper is applied.

B.5 An invertible post-noise transformation

The scalar map $\phi(t) = t + t^3/4$ has derivative $1 + 3t^2/4 > 0$ and tends to the corresponding infinities at both ends of the real line. It is a continuous bijection with continuous inverse. Its coordinatewise extension Φ is therefore a Borel isomorphism. Let A_M be a declared invertible canonicalization of the already noisy observation. Measurability of the maps and their inverses gives

$$\sigma(M, \Phi(A_M O)) = \sigma(M, A_M O) = \sigma(M, O). \tag{26}$$

Both the conditional mean as a random variable and the posterior variance are unchanged. In the displayed feature coordinates, however, the same posterior mean is generally nonlinear. Ridge on transformed features need not attain it; inverse-transform-plus-ridge is a valid control for extracting the preserved information. Applying the transform before the observation noise would define a different channel and is not justified by this argument. Neither is hiding the receiver-dependent map from the decoder. Exact sigma-field equality concerns real-valued observations, whereas float32 rounding is an additional numerical map. The recorded inverse-transform check uses float64 evaluation features and does not establish exact inversion of float32 PFN inputs.

C PRECISE COMPARISON WITH PRIOR MODELS

C.1 EMBEDDING IN WIRELESS LINEAR COMPUTATION BROADCAST

Use the notation of Tan & Jafar (2026), equations (1)–(15), and the following substitution:

Their parameter	Present specialization
K, source dimension	n, n
d_i, d_i'	$1, 1$
V_i, V_i'	$e_{\pi(i)}, e_i$
M, N_i, T	$1, 1, r$
H_i, E	$I_r, B^\top$
$\sigma_{c,i}^2, \sigma_{s,i}^2$	ν, σ^2
ω_i, P_{avg}	$1, P/r$

For their canonical $P_{\text{avg}} = 1$, rescale $E = \sqrt{r/P}\, B^\top$, the broadcast by $\sqrt{r/P}$, and channel-noise variance to $r\nu/P$. Their Proposition 2 gives fixed-precoder MMSE, Proposition 4 fixes decoders, and Proposition 5 establishes stationarity. None is a global paired-family characterization.

The following clarifications follow directly from our objective. Shared versus receiver-specific channel noise does not distinguish expected sum MSE: only each receiver's marginal target/observation law enters. For perfect local observations, use unwhitened conditional covariance; a whitening formula dividing by σ cannot be substituted at zero. Zero encoder and zero decoders are a coordinatewise optimum with loss n, whereas a nonzero pair code has lower loss. Thus a stationary-point guarantee cannot replace the global converse or its equality analysis.

C.2 WHAT THE OLDER OPTIMIZATION RESULTS SUPPLY

Goela & Gastpar (2012, Theorem 3, equation (43)) solve a noisy-network encoder quadratic program conditional on a fixed decoder. Their Theorem 5, equation (52), relaxes a rank constraint. Neither step by itself rules out all cross-pair mixtures under our shared encoder. Wang et al. (2013, Theorems 1–2) analyze limiting signal-to-noise regimes of a relay problem: its noisy uplink and receiver-specific downlinks constrain the effective broadcast. These comparisons acknowledge earlier paired estimation while separating it from a finite-noise all-optimizer statement.

An exact predecessor for the no-local boundary. Schizas et al. (2007, Theorem 4, equations (33)–(36)) solve the noisy single-sensor compression–estimation problem globally. In their notation, take target $s = Jx$, observation covariance $\Sigma_{xx} = I_n$, compression $C = B \in \mathbb{R}^{r\times n}$, channel $D = I_r$, and noise covariance $\Sigma_{zz} = \nu I_r$. Their power constraint is then exactly $\operatorname{tr}(CC^\top) \le P$. The source relevance eigenvalues are all 1 and the channel precision eigenvalues all $1/\nu$; their allocation sets every active squared gain to P/r. Their MMSE is therefore

$$n - \sum_{i=1}^{r} \frac{P/r}{\nu + P/r} = n - \frac{rP}{P + r\nu}.$$

This is established overlap, not a new no-local-information theorem. In the Gaussian specialization, their optimal linear decoder is also the measurable MMSE decoder. Their Theorem 5, equations (38)–(41), handles correlated multisensor data by optimizing one sensor while fixing the others; that result should not be confused with their global single-sensor solution.

Our informative-locality objective instead sums risks for different permitted observations (u_v, z). Giving one fusion center every u_v changes this information restriction: with exact locality, the center already knows the whole source and has zero risk. Independent single-receiver optimizations also do not enforce one common encoder B. Thus the cited theorems do not supply the paired-entry converse, its all-Gram equality conditions, or the fixed invariant-Gram optimum.

Zero-delay Gaussian source–channel coding. Floor et al. (2013, Section 2, equation (3); Section 3.2, equations (15)–(18)) consider M source components, M orthogonal noisy channel outputs,

and one receiver reconstructing every component without private receiver-specific observations. Their cooperative linear scheme decorrelates the source and allocates power among its modes. For unit-variance white sources and total power $P = MP_{\text{there}}$, their per-component value becomes $\nu/(P/M + \nu)$, the dimension-matched classical no-local case. Neither their cooperative nor distributed encoder topology gives the present paired receiver information restriction or its symmetry comparison. Their Proposition 1 uses rate–distortion and channel capacity to give a bound without a zero-delay restriction; it is not an all-Gram zero-delay converse. Their nonlinear mappings in Section 4 act *before* channel noise and can change retained information. Our ϕ acts *after* noise and canonicalization, preserving the ideal observation information. We do not claim a new nonlinear source–channel coding mechanism or superiority over their nonlinear encoders.

A concrete reason to retain fixed-linear scope. For one independent Gaussian pair with exact own-coordinate observations, our $n = 2, r = 1$ optimum gives per-receiver distortion

$$d_{\text{linear}} = \frac{2\nu}{P + 2\nu}.$$

Ordinary Gaussian complementary delivery has rate $\frac{1}{2}\log_2(1/d)$ in the symmetric case (Timo et al., 2010, Example 2, equations (73)–(75)). Combining this established source-coding result with real AWGN common-message capacity $\frac{1}{2}\log_2(1 + P/\nu)$ yields, with increasingly long blocks, distortions approaching

$$d_{\text{block}} = \frac{\nu}{P + \nu} < d_{\text{linear}} \qquad (P > 0).$$

This calculation enlarges the coding class across independent source draws. It does not assert an improvement by a one-shot nonlinear encoder, but it precludes an unrestricted information-theoretic reading of the linear optimum.

C.3 Why symmetrization can leave the feasible representation class

Let G_S be the all-plus optimal Gram selecting r pairs, where S is uniform over the r-subsets of the $n/2$ demand pairs. Each pair is selected with probability $2r/n$, so an elementary average gives

$$\mathbb{E}_S G_S = \frac{P}{n}(I + J), \qquad \operatorname{tr}(\mathbb{E}_S G_S) = P, \qquad \operatorname{rank}(\mathbb{E}_S G_S) = \frac{n}{2}.$$

The same average results from uniformly translating any fixed all-plus selection: the cycle rotations and cube translations act transitively on the demand pairs, so each pair is visited with frequency $2r/n$. More generally, translating a fixed optimum with signs $\varepsilon_1, \ldots, \varepsilon_r$ gives $\overline{G} = (P/n)(I + cJ)$, where $c = r^{-1}\sum_{j=1}^{r} \varepsilon_j$. Its eigenvalues are $(P/n)(1 + c)$ and $(P/n)(1 - c)$, each with multiplicity $n/2$. Thus its rank is $n/2$ for a common sign and n for mixed signs. For $r < n/2$, the average therefore cannot be realized by a fixed r-channel linear encoder. A random realization remains feasible and attains the unrestricted fixed-linear risk, with an equivariant prediction law once selection is shared with the receivers. The selection metadata is assumed available; its communication cost is not included in the real channel count.

This is consistent with probabilistic symmetry theory (Bloem-Reddy & Teh, 2020) and SymPE's risk projection theorem (Lawrence et al., 2025, Theorem 6.1). Averaging predictors in a function space and averaging feasible physical implementations are distinct operations. The group-average predictor would need its own valid factorization through the permitted receiver observations, common broadcast, rank, energy, and isotropic noise. Our calculation provides an explicit failure of closure for the encoder Grams, not a counterexample to those risk projection results.

Attribution boundary. Gaussian conditioning, pair cancellation, probabilistic symmetrization, and information preservation under an invertible observation map are background. The proposed contributions are the paired-family global optimum with every optimal Gram and the matched-resource fixed invariant-Gram optimum, followed by their qualified locality consequences. The comparisons above delimit model and theorem scope; they do not establish exhaustive priority over related coding formulations.

D EMPIRICAL PROTOCOL, IMPLEMENTATION, AND AUDIT

D.1 ADMISSIBLE INFORMATION AND THE IMPLEMENTED WRAPPER

The stochastic protocol uses independent complete Gaussian source/noise draws for context, validation, and query rows. A row is one receiver's permitted view of a fresh complete source vector; it is not another vertex of a source graph already represented in the same context or query table. The implementation uses NumPy pseudorandom generators with separate hash-tagged source, local-noise, and broadcast-noise streams. Phase, replicate, split, graph size, rank, graph type, and receiver role enter the stream key. The invariant role shares the selected role's standardized base stream to support paired comparisons; raw/transformed blocks share draws, and context sizes use nested prefixes. These deterministic simulation conventions reproduce the intended stochastic experiment; distinct hashes are not a mathematical proof of independence. The Gaussian theorem is about the ideal random-draw model.

The evaluation layer stores complete source information only long enough to generate features, labels, and audit quantities. The actual model calls are `fit(X,y)` for context features and context labels, and `predict(X)` for feature-only validation and query calls. Only the regression mean is used. The worker retains query targets, exact conditional means, variances, and stream metadata for scoring, and does not pass these objects to the model. The declared decoder initialization uses a separately tagged reproducibility seed. The mathematical result treats decoder randomness as independent of fresh source/noise draws; it does not model an adversarial program reconstructing pseudorandom samples from archive metadata. Recorded data-generation seeds are reproducibility information and are excluded from the inference interface.

The adapter supplies feature-only query chunks of 128 rows to `predict`, so the interface permits dependence on the other features in a chunk. The retained chunk-size probes demonstrate that predictions can change with batching, but do not identify architectural query attention as the cause. No rowwise batch-invariance assumption is needed for the conditional-risk identity; shared fitted contexts already preclude treating query losses as independent replicate estimates.

Receiver identity, pair selection, channel assignment, and the known encoder are admissible input-independent metadata for the wrapper. These are used to canonicalize features rather than appended as query source information. For the invariant encoder, $BB^\top = (P/r)I$. The implementation uses the cycle permutation convention $(T_v x)_j = x_{j+v \bmod n}$ and

$$C_v = \frac{r}{P} B T_v B^\top, \qquad C_v B = B T_v, \qquad C_v C_v^\top = I.$$

It supplies $(u_v, C_v z)$, or $C_v z$ without local information. This is the inverse-translation convention for the usual group action $T_g e_j = e_{gj}$: it maps receiver v's joint feature/target law to receiver 0's law. For a selected paired receiver, the wrapper swaps its active broadcast channel into position 0; unselected receivers form the other stratum. The same rule applies to every context and query row. Weights are $2r/n$ and $1 - 2r/n$; at $r = n/2$ the unselected stratum is absent. The rule concerns mean marginal risk and does not reconstruct the simultaneous error covariance of all receivers sharing one realized broadcast.

The operation order is source draw, linear broadcast plus independent noise, metadata canonicalization, optional coordinatewise ϕ, then model input conversion. For $\phi(t) = t + t^3/4$, the real inverse is

$$\phi^{-1}(s) = \frac{4}{\sqrt{3}} \sinh\left[\frac{1}{3} \operatorname{arsinh}\left(\frac{3\sqrt{3}}{4} s\right)\right].$$

The Gaussian information floor is unchanged before numerical rounding. Float32 conversion or internal model preprocessing may discard information; they cannot invalidate the original lower bound, but need not preserve its attainability. The inverse-ridge check uses the float64 evaluation features and does not establish exact bit-level inversion of PFN inputs.

D.2 STATISTICS AND THE ORIGINAL FIVE-REPLICATE RESULTS

For row i, let $\delta_i = \widehat{y}_i - \mu_i$ and $\epsilon_i = y_i - \mu_i$. The observed-batch identity is

$$\widehat{\mathrm{MSE}} = \frac{1}{N} \sum_i \epsilon_i^2 + \frac{1}{N} \sum_i \delta_i^2 - \frac{2}{N} \sum_i \delta_i \epsilon_i.$$

The floor-plus-excess quantity $\tau + N^{-1}\sum_i \delta_i^2$ is the conditional expected batch MSE in the ideal protocol. Its position above τ is true by construction, so it is not an independent empirical verification of a lower bound. All identities extend with the declared deterministic receiver weights. The primary raw contrast is the paired within-replicate difference between invariant and paired actual MSE. For the original five replicate values $a_1, \ldots, a_5$, reported intervals are $\bar{a} \pm t_{0.975,4} s_a/\sqrt{5}$, with $t_{0.975,4} = 2.776445105\ldots$. They are nominal small-sample intervals; normality and simultaneous multiplicity-adjusted coverage are not asserted. Rowwise iid standard errors are diagnostic only because contexts and feature-only transduction can couple predictions across rows.

Input / decoder	Paired excess [95% interval]	Invariant excess [95% interval]
Raw / ridge	.003702 [.002233, .005172]	.006384 [.004802, .007967]
Raw / TabPFN	.001551 [.000682, .002419]	.007511 [.004288, .010734]
Raw / TabICL	.002425 [.000992, .003858]	.006151 [.003425, .008878]
ϕ / ridge	.086577 [.077013, .096141]	.059178 [.053202, .065154]
ϕ / TabPFN	.002200 [.001553, .002847]	.010008 [.006936, .013080]
ϕ / TabICL	.005096 [.003445, .006748]	.013277 [.008617, .017937]
ϕ / inverse-ridge	.003702 [.002233, .005172]	.006384 [.004802, .007967]

Table 2: Original-cohort weighted excess at context $1{,}024$. Five independent context/query replicates; intervals are nominal, with no multiplicity adjustment.

Original-cohort actual raw paired/invariant MSE means are $0.565252/0.709521$ for ridge, $0.563176/0.710400$ for TabPFN, and $0.565166/0.709857$ for TabICL.

The ridge actual gap is 0.144269 $[0.111315, 0.177223]$. The transformed-minus-raw excess increases, computed as paired differences, are $0.082875/0.052794$ for ridge, $0.000649/0.002496$ for TabPFN, and $0.002671/0.007126$ for TabICL (paired/invariant). The transformed-only names `raw_ridge` and `inverse_ridge` create four structurally unavailable rows in the automatic same-name raw/transformed contrast table. They are not missing experiments: their proper reference is raw-block `ridge`, compared directly by the auditor. Maximum inverse-control prediction discrepancy is 5.1071×10^{-15}.

D.3 Prospectively specified replication extension

The extension was specified after seeing the original five replicates and receiving automated feedback. It is a prospective replication extension, not part of the original confirmatory plan or a formal public preregistration. The frozen umbrella design specifies 25 fresh cycle replicates (E1), a resource-selected cube cohort (E2), the deterministic affine calculation (E3), and bounded repeatability diagnostics (E4). An engineering profile with distinct streams selected ten cube replicates before the scientific job list was frozen or its outcomes generated. The plan, source, job list, seed keys, pre-outcome selection, attempts, and runtime records accompany the supplement. No outcome-dependent stopping, replacement seeds, decoder tuning, or extra learned-encoder starts were used.

The new cycle design has $1{,}050$ receiver jobs: 700 PFN fits and 350 baseline jobs. The cube has 180 receiver jobs: 120 PFN fits and 60 baseline jobs. All $1{,}230$ jobs completed; their audit reconstructs $2{,}260$ receiver records, including the additional controls scored within baseline jobs. The complete independently generated context/query replicate remains the sampling unit. Paired arms and presentations share their declared base streams, and the raw context sizes remain nested. The original five are retained separately; the cycle extension has 25 new replicates and the cube has ten. Neither diagnostics nor query rows increase these counts.

For k replicate aggregates or paired differences, the nominal interval is $\bar{a} \pm t_{0.975,k-1} s_a/\sqrt{k}$, with $k = 25$ for E1 and $k = 10$ for E2. The frozen primary family contains only the two raw-context-$1{,}024$ PFN invariant-minus-paired actual-MSE contrasts in E1. Each uses $t_{0.9875,24}$, giving Bonferroni familywise nominal 95% coverage under the Student-t assumptions. This multiplicity adjustment does not make a finite-sample nonparametric guarantee. Transformation, controls, cube, component analyses, and threshold attainment are secondary; their 95% intervals are descriptive and not multiplicity-adjusted. The frozen 0.02 excess level is not a simultaneous hypothesis test.

Component differences are formed within replicate before any averaging. The paired MSE difference is exactly the sum of oracle-batch, excess, and residual cross-term differences. The variance of

Decoder	MSE gap: familywise 95%	Gap minus 16/117: nominal 95%
TabPFN	$0.140587\ [0.129777, 0.151398]$	$0.003835\ [-0.005497, 0.013167]$
TabICL	$0.137919\ [0.127279, 0.148559]$	$0.001167\ [-0.008017, 0.010351]$

Table 3: The two prospectively specified raw-cycle contrasts (25 new replicates, df=24), each invariant minus paired. The MSE intervals use Bonferroni adjustment (97.5% individually); deviation-from-floor-gap intervals are descriptive, unadjusted.

Block / decoder	Paired excess	Invariant excess
raw / Ridge	$0.00529\ [0.00430, 0.00627]$	$0.00675\ [0.00578, 0.00772]$
raw / TabPFN	$0.00216\ [0.00171, 0.00260]$	$0.00770\ [0.00685, 0.00855]$
raw / TabICL	$0.00346\ [0.00256, 0.00437]$	$0.00660\ [0.00578, 0.00743]$
nonlinear / Ridge	$0.09595\ [0.09138, 0.10052]$	$0.06120\ [0.05890, 0.06350]$
nonlinear / TabPFN	$0.00317\ [0.00265, 0.00368]$	$0.01062\ [0.00973, 0.01151]$
nonlinear / TabICL	$0.00616\ [0.00528, 0.00705]$	$0.01511\ [0.01410, 0.01613]$
nonlinear / Inverse + ridge	$0.00529\ [0.00430, 0.00627]$	$0.00675\ [0.00578, 0.00772]$
no_local / Ridge	$0.00678\ [0.00580, 0.00775]$	$0.00669\ [0.00575, 0.00762]$
no_local / TabPFN	$0.00210\ [0.00162, 0.00258]$	$0.00729\ [0.00625, 0.00833]$
no_local / TabICL	$0.00318\ [0.00233, 0.00402]$	$0.00534\ [0.00454, 0.00613]$
endpoint / Ridge	$0.00324\ [0.00274, 0.00375]$	$0.00333\ [0.00284, 0.00382]$
endpoint / TabPFN	$0.00511\ [0.00417, 0.00605]$	$0.00960\ [0.00888, 0.01033]$
endpoint / TabICL	$0.00505\ [0.00446, 0.00565]$	$0.00503\ [0.00444, 0.00562]$

Table 4: New cycle cohort (25 replicates, df=24), 1,024 context rows. Mean sampled excess and nominal 95% replicate intervals. Inverse ridge uses the float64 evaluation pipeline.

this sum retains their covariances; it is not the sum of marginal variances or interval widths. The supplement records the full component covariance matrices, intervals for the endpoint oracle-batch term, and intervals for raw MSE gap minus the exact information gap. Equality of population floors constrains expected oracle error; it does not force a particular held-out batch or restricted fitted decoder to have zero arm difference.

The original TabPFN endpoint gap is 0.014199, versus 0.000461 in E1; TabICL changes from 0.008630 to -0.004145. Their original shared oracle-batch component is 0.006945, versus -0.003554 in E1. The E1 TabICL endpoint excess difference is -0.000026 and its cross-term difference is -0.000565, so its negative MSE difference is mostly the sampled oracle-batch contribution. These comparisons are retained separately rather than obscured by a pooled endpoint average. The raw primary gaps, in contrast, remain close to the exact information gap.

D.4 Context-512 excess disclosure

The frozen nested raw-context arm is also shown in Figure 1. The following post-review descriptive breakdown distinguishes group intervals from individual replicate aggregates; it changes neither the primary family nor the frozen 0.02 diagnostic. Seven of the 100 new model/encoder/replicate aggregates exceed 0.02: invariant TabICL replicates $1003, 1008, 1021, 1024$ and invariant TabPFN replicates $1008, 1021, 1024$. The maximum is 0.030512 (TabPFN, 1008). All four group means and nominal upper limits remain below 0.02; their maxima are 0.014984 and 0.017178, respectively. The original cohort has two of 20 exceedances, both invariant replicate 1, for TabPFN and TabICL; its maximum is 0.029781 (TabICL). Its maximum group mean is 0.015400, but its maximum nominal upper limit is 0.026525. Thus the group-level and individual-aggregate diagnostics differ, as do the two cohorts. All observations are retained.

D.5 Separate Boolean-cube implementation study

The Q_5 cohort has 32 vertices, pairing $v \mapsto v \text{ xor } 31$, $r = P = 8$, broadcast-noise variance $1/16$, and exact local observations. Every replicate uses $1{,}024$ context rows and $2{,}048$ query rows, with raw and transformed presentations. The invariant encoder uses the first eight even-Hamming-weight Walsh characters in ascending integer order, $0, 3, 5, 6, 9, 10, 12, 15$, with signs positive at vertex zero. Channel ordering, selected-pair ordering, and the metadata wrapper were fixed before outcomes. The XOR canonicalization acts by the corresponding diagonal character signs. Independent deterministic

Paired quantity	Mean [nominal 95% interval]
Endpoint / TabPFN / MSE	$0.000461\ [-0.003016, 0.003938]$
Endpoint / TabPFN / oracle batch	$-0.003554\ [-0.006982, -0.000127]$
Endpoint / TabPFN / decoder excess	$0.004492\ [0.003447, 0.005538]$
Endpoint / TabPFN / cross term	$-0.000477\ [-0.001321, 0.000367]$
Raw / TabPFN / gap minus $16/117$	$0.003835\ [-0.005497, 0.013167]$
Endpoint / TabICL / MSE	$-0.004145\ [-0.007685, -0.000605]$
Endpoint / TabICL / oracle batch	$-0.003554\ [-0.006982, -0.000127]$
Endpoint / TabICL / decoder excess	$-0.000026\ [-0.000684, 0.000632]$
Endpoint / TabICL / cross term	$-0.000565\ [-0.001504, 0.000375]$
Raw / TabICL / gap minus $16/117$	$0.001167\ [-0.008017, 0.010351]$

Table 5: New cycle paired decomposition (25 replicates, df=24). Every difference is formed within replicate after fixed receiver weighting. The endpoint exact information gap is zero. Component covariance is retained in the machine tables; marginal interval endpoints are not combined.

Cohort	Decoder	Paired excess [nominal 95%]	Invariant excess [nominal 95%]
New 25	TabPFN	$0.004473\ [0.003394, 0.005552]$	$0.014984\ [0.012791, 0.017178]$
New 25	TabICL	$0.007026\ [0.005537, 0.008516]$	$0.014229\ [0.012042, 0.016416]$
Original 5	TabPFN	$0.003583\ [0.000701, 0.006465]$	$0.015110\ [0.005555, 0.024665]$
Original 5	TabICL	$0.006250\ [0.003953, 0.008547]$	$0.015400\ [0.004274, 0.026525]$

Table 6: Raw inputs with 512 context rows. Means and nominal replicate Student-t intervals use 24 and 4 degrees of freedom for the new and original cohorts, respectively. No pooling or multiplicity adjustment.

covariance calculations verify every vertex's joint feature–target law, receiver weights, energy, row Gram, and posterior variance, including an additional $r \neq P$ case.

Block / decoder	Paired excess	Invariant excess
raw / Ridge	$0.00490\ [0.00289, 0.00691]$	$0.00563\ [0.00333, 0.00794]$
raw / TabPFN	$0.00184\ [0.00129, 0.00239]$	$0.00868\ [0.00576, 0.01159]$
raw / TabICL	$0.00316\ [0.00143, 0.00488]$	$0.00739\ [0.00502, 0.00976]$
nonlinear / Ridge	$0.09583\ [0.09179, 0.09986]$	$0.05997\ [0.05556, 0.06438]$
nonlinear / TabPFN	$0.00292\ [0.00237, 0.00346]$	$0.01290\ [0.01004, 0.01577]$
nonlinear / TabICL	$0.00564\ [0.00394, 0.00734]$	$0.01581\ [0.01333, 0.01830]$
nonlinear / Inverse + ridge	$0.00490\ [0.00289, 0.00691]$	$0.00563\ [0.00333, 0.00794]$

Table 7: Separate Q_5 cohort (10 replicates, df=9), 1,024 context rows. Mean sampled excess and nominal 95% intervals. Fixed Walsh basis and XOR wrapper; no pooling with cycles.

A post-review decomposition on the same retained query batches gives an oracle invariant-minus-paired MSE contrast of $0.152153\ [0.129944, 0.174361]$ (nominal 95%, 9 degrees of freedom), common to the raw and transformed presentations and all three decoders. Its deviation from $16/117$ is 0.015400, accounting arithmetically for 99.1%, 73.9%, and 84.3% of the raw Ridge, TabPFN, and TabICL gap deviations, respectively. The remaining contrasts are 0.000141, 0.005443, and 0.002877. Each remainder includes both sampled excess and the realized residual cross term; it is not pure decoder excess. These shares are descriptive arithmetic, not a causal explanation or a variance decomposition, and do not extend uniformly to transformed inputs.

All eight cube PFN model/arm/presentation group means and nominal upper excess limits are below 0.02. One of the 80 replicate aggregates exceeds it: transformed-invariant TabICL replicate 2010, with excess 0.021086. No observation is excluded. At the principal setting, the independent-row canonical raw feature–target marginals for the cycle and cube agree up to latent orthogonal coordinates. For invariant encoders, let b_C and b_Q be their receiver-zero columns and set $u = b_C - b_Q$. The Householder map $H = I - 2uu^\top/(u^\top u)$ maps b_C to b_Q and preserves isotropic broadcast covariance; the full mapped joint-covariance discrepancy is below 8.89×10^{-16} in the independent deterministic check. Paired selected-receiver covariance agrees exactly. These are marginal row laws, not an assertion about simultaneous errors of all receivers. The practical decoder need not respect such rotations, and a coordinatewise cubic map does not generally commute with them. The cohort

Presentation / decoder	MSE gap	Gap minus $16/117$
raw / Ridge	$0.152294\ [0.131923, 0.172664]$	$0.015542\ [-0.004829, 0.035912]$
raw / TabPFN	$0.157596\ [0.136087, 0.179104]$	$0.020843\ [-0.000665, 0.042352]$
raw / TabICL	$0.155029\ [0.134106, 0.175953]$	$0.018277\ [-0.002646, 0.039201]$
nonlinear / Ridge	$0.113878\ [0.091045, 0.136711]$	$-0.022874\ [-0.045708, -0.000041]$
nonlinear / TabPFN	$0.160105\ [0.137931, 0.182279]$	$0.023353\ [0.001179, 0.045527]$
nonlinear / TabICL	$0.159062\ [0.137557, 0.180567]$	$0.022310\ [0.000805, 0.043815]$

Table 8: Separate cube invariant-minus-paired differences (10 replicates, df=9). All intervals are nominal descriptive 95%. These finite fitted-decoder outcomes do not identify a causal effect of graph topology.

is therefore a check of the stated cube basis and wrapper, not an isolated test of graph topology or robustness to realistic graph/source distributions. It is never pooled with the cycle.

D.6 Population affine diagnostic for transformed features

This deterministic calculation was added after inspecting the original ridge results. It isolates the best affine predictor from the displayed features, separately from finite-context estimation and regularization. Let X be a centered Gaussian vector of permitted features after noise and canonicalization, with $K = \mathrm{Cov}(X)$ and $c = \mathrm{Cov}(X, Y)$. Apply $\phi(t) = t + at^3$ to every feature, including the local observation when present, as in the implementation. For $a = 1/4$, set $D = \mathrm{diag}(1 + 3aK_{ii})$. Gaussian moments give

$$\mathbb{E}[X_i^3 Y] = 3K_{ii}c_i, \qquad \mathbb{E}[X_i^3 X_j^3] = 9K_{ii}K_{jj}K_{ij} + 6K_{ij}^3.$$

Expanding the transformed products therefore yields

$$\mathrm{Cov}(\phi(X), Y) = Dc, \qquad \Sigma_\phi := \mathrm{Cov}(\phi(X)) = DKD + 6a^2 K^{\circ 3},$$

where $K^{\circ 3}$ is the entrywise cube. Since $K \succ 0$, $D \succ 0$, and the Schur product theorem gives $K^{\circ 3} \succeq 0$, $\Sigma_\phi \succ 0$. The optimal affine intercept is zero by centering; its population risk is

$$R_{\mathrm{aff},\phi} = \mathrm{Var}(Y) - (Dc)^\top \Sigma_\phi^{-1} (Dc).$$

We evaluate this expression by covariance solves for each receiver stratum, then use the declared receiver weights. For the principal C_{32} setting, the paired selected and unselected strata have weights $1/2$ each; the unselected target is independent of its features and has risk 1. The calculation uses each encoder's declared native channel basis after metadata canonicalization. A coordinatewise nonlinear transform is basis dependent; equality of raw Gaussian Bayes risks under latent orthogonal changes does not make this affine diagnostic basis independent.

Representation	Bayes floor	Transformed affine risk	Affine excess
Paired	0.555556	0.643697	0.088141
Invariant	0.692308	0.743787	0.051479

Table 9: Population affine diagnostic for the principal C_{32} setting, $r = P = 8$, $\nu = 1/16$, with exact local observations. Values are receiver-weighted population calculations, not fitted ridge estimates or new model replicates.

The paired affine excess is larger, although its total affine risk is still smaller. Equal broadcast marginals do not equalize the joint feature–target structure or the cost of restricting the decoder to affine functions after the nonlinear map. The $a = 0$ calculation agrees with the Gaussian oracle within 2.23×10^{-16}. Independent four-node Gauss–Hermite integration of the degree-at-most-six moment polynomials agrees within 2.85×10^{-14} for feature covariances and 4.45×10^{-16} for risks across the audited settings. No model inference or new random draws enter this calculation. It concerns ideal real-valued transformed features, not rounded PFN inputs. Sampled ridge excess can lie below this population affine excess through query-sampling fluctuation; the calculation does not predict the exact finite-context ridge error.

D.7 EXACT CONTROLS AND FITTED ENCODERS

The no-local control has rank 8, power 8, and exact floor $13/17$ for both encoders. The endpoint has rank 16, unchanged power 8, and floor $1/5$. In the original five replicates, TabPFN has an endpoint MSE gap 0.014199 $[0.005734, 0.022665]$; its exact sample decomposition is $0.006945 + 0.005108 + 0.002146$ (oracle-batch, excess, cross-term differences). The oracle-batch difference has expectation zero under the ideal protocol; its paired sampling uncertainty depends on the cross-encoder residual covariance. Table 10 now reports that uncertainty from the original five replicates, along with deviations of the raw MSE gaps from the exact information gap. Each component difference is formed within replicate before computing its interval; the components are not treated as independent, and the interval for their sum uses their joint pairing rather than adding marginal intervals. TabICL's corresponding excess difference is 0.000328. These are finite-decoder observations, not evidence against the equal-floor theorem. Neither nonsignificant gaps nor small excess establish equivalence. The same exact-zero oracle expectation holds for E1. Independent reconstruction of its retained endpoint targets and conditional means agrees with the archived arrays to numerical precision and recovers both $1/5$ floors. Its oracle-batch contrast has nominal $t = -2.140$ on 24 degrees of freedom; this is one of many secondary descriptive comparisons, not an additional multiplicity-adjusted primary test.

Contrast	Mean	Nominal 95% interval
Endpoint / oracle batch	0.006945	$[-0.002862, 0.016751]$
Endpoint / TabPFN excess	0.005108	$[0.001524, 0.008692]$
Endpoint / TabPFN cross term	0.002146	$[-0.001130, 0.005422]$
Endpoint / TabPFN actual MSE	0.014199	$[0.005734, 0.022665]$
Raw / TabPFN gap minus $16/117$	0.010472	$[-0.023036, 0.043980]$
Raw / TabICL gap minus $16/117$	0.007939	$[-0.022976, 0.038853]$

Table 10: Reanalysis of the original five C_{32} replicates at 1,024 context rows. Every contrast is invariant minus paired. Intervals use paired replicate differences with four degrees of freedom; they are descriptive and not multiplicity-adjusted. The endpoint population information gap is zero. The oracle-batch component is common to the evaluated decoders.

The CPU audit contains 492 exact-grid records and 66 control records. A depth-one case with $n = 32$, $r = P = 4$ gives paired feasible risk 0.777778 and invariant lower bound 0.818182; this is a separation certificate, not an exact positive-depth optimum. Outside the theorem's white-source domain, source correlation 0.9 with actual covariance-weighted power matched at 16 gives compared fixed-code risks 0.544114 (paired) and 0.446012 (invariant). For the pure heat target the corresponding fixed-code risks are 0.076916 and 0.070277. These reversals mark the theorem's scope and do not identify optimal encoders for the altered problems.

The learned-encoder experiments comprise three population starts at each local-noise variance 0 and 1, and three separately sample-trained runs at variance 0. The population optima are respectively $5/9$ and $9/13$. All six fits attain the relevant optimum within 7×10^{-14}, with eight active pair blocks. The sample-trained runs' validation-selected steps are 5,250, 5,750, and 6,000; their fixed-encoder Bayes risks are 0.556085, 0.556063, and 0.556076. Independent reconstruction verifies saved initializations, fitted weights, validation selection, and held-out predictions and targets. Three starts provide finite experimental evidence of structural recovery, not a guarantee for other initializations, tasks, or optimizers.

D.8 FROZEN PROVENANCE AND NUMERICAL CHECKS

The confirmatory configuration uses seed 3961491, fresh `confirmatory` streams, and was frozen before the final-query draws. All 221 jobs completed on their first attempt: 140 PFN fits (70 per model), 70 baseline receiver jobs, two CPU audit jobs, and nine learned encoder fits. There are 380 receiver records, 270 weighted replicate aggregates, 54 weighted summary groups, and 57 populated paired comparisons. Earlier development runs informed protocol choice and are not additional confirmatory replicates.

The original evidence was consolidated for anonymous export. Its numerical audit verifies archive integrity, frozen source and protocol identities, job/output hashes, independent reconstruction of

Gaussian observations, targets, conditional means and variances, split isolation, paired streams, nested prefixes, all saved metrics, ridge predictions, and aggregate statistics. The largest discrepancy is below 7.57×10^{-14} for independently solved ridge predictions; scalar metric/interval discrepancies are below 4.45×10^{-16}. Pretrained checkpoint bytes are not included in the export: recorded pinned identities are checked by the external auditor, while the original worker verifies local checkpoint bytes before inference. The original source and audit code are retained for reproduction.

To resolve the proof review's missing implementation evidence, we additionally ran ten targeted existing offline checks, including all-vertex joint-law canonicalization tests on the 16-cycle and 16-vertex cube, and checks of reference risks, feasible invariant ranks, stream isolation, and the nonlinear inverse. A separate deterministic audit checks every vertex in both families for all four frozen 32-cycle blocks (256 vertex/family/block cases). It independently constructs the full joint covariance of canonical raw features and target. Maximum discrepancies from the representative law are 1.56×10^{-15} for covariance, 4.45×10^{-16} for posterior variance, and 2.23×10^{-16} for the weighted mean floor. All 16 frozen source-file hashes match. These checks require no real PFN inference or new scientific draws. The feature-only interface is established by inspected worker/adapter calls; this is not a security claim against arbitrary malicious model code.

D.9 Original repeatability and compute

In the original run, 20 of 70 TabPFN same-chunk validation-repeat probes had nonzero maxima: 15 were at most 2.98×10^{-5} and five were 0.007866, 0.029177, 0.021825, 0.034112, and 0.097784. Each probe predicts the same 16 validation-feature rows in chunks 16, 8, and 16, before the 128-row main-query computation. All 70 TabICL same-chunk maxima were zero, with maximum different-chunk discrepancy 6.47×10^{-6}. We retained the original outputs and replayed the five anomalous TabPFN cells once each with the same frozen jobs and runtime. Independent comparison against original archived references finds every saved array byte-identical, including all $10{,}240$ main-query predictions, with zero recomputed metric change. The replay archive has 65 manifested payload files. Replay logs report zero same-chunk probe maxima and at most 2.14×10^{-5} for different chunks; raw validation-probe arrays were not exported, so those maxima cannot be independently recomputed. The original anomaly's cause remains unknown. The five replays establish repeatability of those query computations in the recorded environment and add zero independent scientific replicates. The same-chunk comparison is between the first and third calls, separated by a changed-chunk call. It therefore does not isolate numerical threading effects from model state, caching, or call-order effects. The saved scalar maxima alone cannot identify any such cause; a CPU result does not establish a GPU-atomic-operation explanation.

The environment was Python 3.12.3, NumPy 2.5.3, SciPy 1.18.1, scikit-learn 1.9.1, CPU PyTorch 2.14.0, TabPFN 8.5.0, and TabICL 2.2.0. The machine reports 22 logical CPUs and 30.79 GiB host RAM; numerical workers use six threads, one ensemble member, and float32 PFN inputs/inference with AMP disabled. Gaussian auditing and ridge use float64. TabPFN uses the pinned default version-3 regressor checkpoint; TabICL uses the pinned version-2 regression checkpoint. Source/checkpoint identities and constructor/resolved configuration records accompany the artifact. Model weights are frozen; no fine-tuning or test-dependent hyperparameter selection is performed.

The original runner took $2{,}483.58$ seconds (41.39 minutes), with peak polled process-tree-plus-supervisor RSS $1{,}073.60$ MiB (1.048 GiB). TabPFN and TabICL workers account for $1{,}451.91$ and 913.60 seconds respectively. All cells completed without resource termination. Memory samples are 0.25 seconds apart; they may miss brief peaks and double-count shared pages, and are not total-machine RAM measurements. There were 55 TabPFN CPU-context-size warnings, without job failure. The supplemental five-cell replay took 106.27 seconds. Original confirmatory predictions and intervals remain the evidence for the historical study; its technical replays do not replace scientific observations. The new cycle cohort supplies the primary prospective replication estimates.

D.10 Extension repeatability, compatibility, and compute

The extension captures the actual permitted feature arrays and predictions from the existing validation call sequence: the same 16 rows in chunks 16, 8, and 16, followed by main-query chunks of 128. Passive capture adds no model calls or random draws. Engineering capture-on/off comparisons preserve model outputs exactly; all normal scientific inference settings remain unchanged. Recorded

RNG-state hashes before and after the calls are unchanged within all 820 scientific PFN fits, but those hashes do not describe every possible internal cache or model state.

Among the 410 scientific TabPFN fits, 92 first-versus-third same-chunk comparisons differ: $87/350$ in E1 (maximum 0.027735) and $5/60$ in E2 (maximum 0.059510). Every TabPFN changed-chunk comparison differs, with maximum 0.194722 across the scientific fits. All 410 TabICL same-chunk comparisons are exact; its changed-chunk maximum is 8.89×10^{-6}. The supplement retains actual probe arrays, hashes, and call order, enabling these maxima to be recomputed. Same-chunk comparisons are separated by a changed-chunk call; this protocol does not isolate threading, cache, or call-order mechanisms.

E4 is a separate bounded diagnostic cohort. For each model, two captured fresh-process executions and a capture-off execution of the matched case produce byte-identical 16-row main predictions. A joint change to one thread and a deterministic setting changes main predictions by at most 1.40×10^{-5} (TabPFN) and 2.99×10^{-6} (TabICL), with both bounds conservatively rounded upward. These small cases do not reproduce every scientific anomaly or identify which component of the joint setting change caused a difference. Engineering/profile probes and E4 are not scientific replicates. The original anomalies remain historically unresolved because their raw probe arrays were not archived; new capture does not retroactively resolve them. The evidence establishes only the reported comparisons in this runtime, not general query-batch invariance.

The original and E1 jobs use matching checkpoint hashes, package versions, explicit constructor settings, data generation, preprocessing, and inference protocol. The data, mathematical, and model modules are byte-identical; worker additions record passive capture/provenance and an E4-only settings switch. Apparent differences in the recorded TabICL configuration representation are process-specific object addresses. Matching this opaque representation does not establish a complete serialization of unreported internal model state. These checks support descriptive pooling of the five original and 25 new cycle replicates under the documented protocol; the supplement includes such 30-replicate summaries as secondary results. The paper's primary estimates use E1 alone, retain the original study separately, and explicitly report the endpoint disagreement.

The extension's charged workflow time is $18{,}110.84$ seconds (5.031 hours), including profiling, E4, and orchestration, and excluding final export. Peak sampled process-tree-plus-supervisor RSS is 1.232 GiB. The recorded pre-export stages remain within the six-hour and 20 GiB limits, with six-thread scientific workers and one supervised model worker at a time. These sampled memory measurements are neither instantaneous maxima nor total-machine RAM use. The anonymous supplement preserves failure/attempt records, frozen source, plans, scientific predictions, and actual probe outputs; pretrained weights are excluded. Scores and manuscript analyses can be audited offline without fetching model weights.

E SUPPLEMENTARY EXPERIMENTAL FIGURES

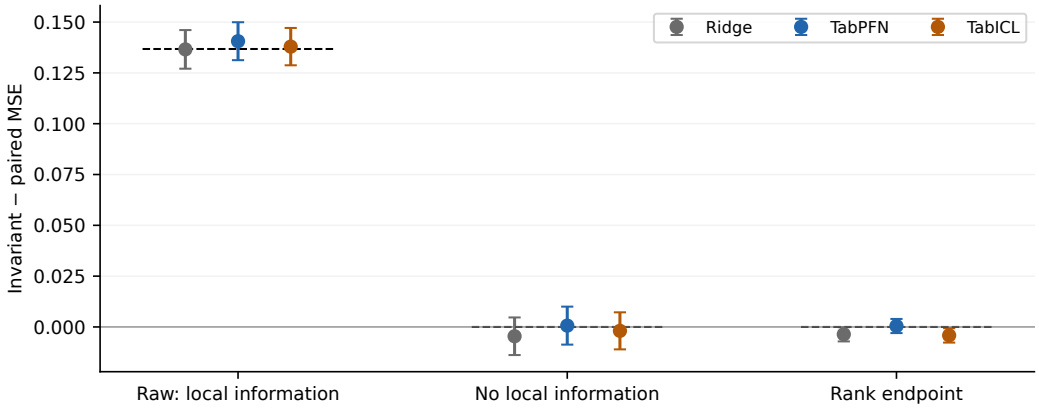


Figure 3: **Mechanism controls with qualified finite-decoder conclusions.** Invariant-minus-paired held-out MSE gaps in the 25 new cycle replicates at 1,024 context rows. Bars are nominal 95% paired t intervals. The dashed segment marks the principal information gap $16/117$; the horizontal zero line marks the equal-floor controls. The no-local block has floor $13/17$ in both arms. The rank endpoint uses $r = 16$ with unchanged $P = 8$ and floor $1/5$. The new cohort does not reproduce the original positive TabPFN endpoint gap, while TabICL has a small negative gap. Equality of information floors does not require equality of fitted-decoder errors.

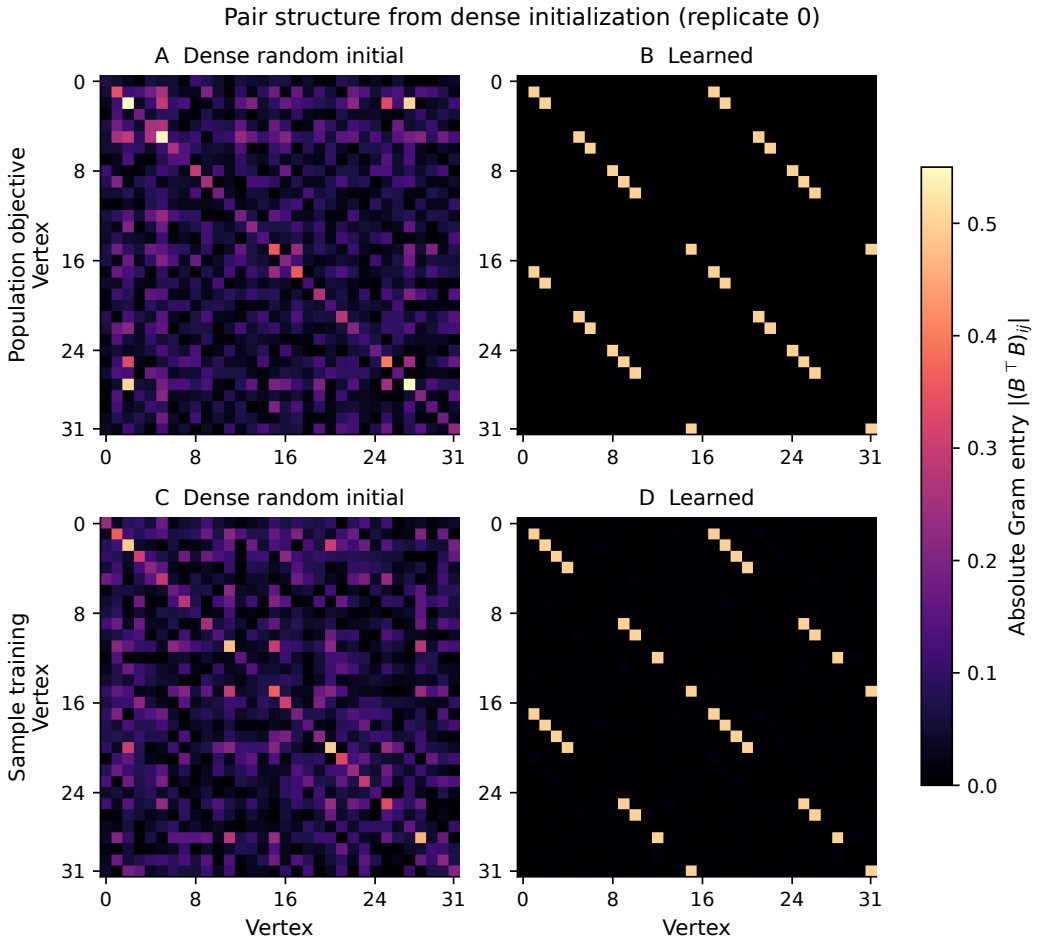


Figure 4: **Learned encoder structure.** Initial and final Grams for prespecified replicate 0 of population-objective optimization and sample training, alongside the pair structure. The dense parameterization does not impose pairing. The population fits attain the optimum numerically; sample-trained encoders retain small off-pair energy. These finite runs support discovery of the predicted structure, rather than an optimizer convergence theorem. All starts are included in the numerical results.