Pitfalls and Remedies for Multi-Task Bayesian Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bayesian optimization (BO) routinely warm-starts target experiments with data from related source tasks, and the multi-task Gaussian process (MTGP) is the textbook surrogate for the job. We revisit this default in a controlled setting and find that it misestimates the cross-task correlation even on the simplest non-trivial case: affinely related source and target tasks, where a working transfer-learning method should obviously succeed. We trace the failure to two independent structural mechanisms. Per-task standardization, the textbook fix for the affine slice ambiguity, propagates a finite-sample alignment error into the recovered correlation. The marginal likelihood itself identifies the correlation only at a per-sample rate that a Gaussian process at non-overlapping designs further dilutes. We propose three conservative remedies that follow from the analysis: promoting per-task means and scales to model parameters, restricting the task covariance to non-negative correlations, and co-locating part of the source and target designs. Across a multi-task BO grid and a transfer-learning sweep on an instruction-following benchmark, these remedies recover the vanilla baseline on the simple instances, while the broader failure persists on harder instances and across most rank-based and latent-context variants.

1 Introduction

BO [Frazier, 2018, Snoek et al., 2012] is a workhorse for sample-efficient experimentation, and in production settings target experiments rarely arrive in isolation. BO transfer learning (BOTL) is the default story whenever a target experiment has a related predecessor: gather the source data, fit an MTGP, and expect fewer target evaluations to find the optimum. Yet on three affinely-related tasks drawn from a standard benchmark function, the textbook MTGP recovers task correlations of the wrong sign. Affinely-related tasks are the textbook example a working transfer-learning method must handle: the source is a perfect linear image of the target, every standard MTGP variant should recover near-perfect correlations, and any reasonable practitioner would expect transfer to help. Affinely-related tasks give a controlled setting in which the achievable performance ceiling is known: an oracle drawing all data from the source, or given the per-task standardization parameters, recovers the target with no transfer cost. We benchmark against this implicit ceiling rather than running the oracle directly. We find the opposite, and similar pathologies persist not only for the textbook Intrinsic Coregionalization Model (ICM) but for nearly every multi-task and rank-based variant in common use. Across our multi-task BO grid it loses to a single-task Gaussian process (GP) on most base functions – the textbook signature of *negative transfer* [Pan and Yang, 2010, Wang et al., 2019]. We trace the failure to a structural identifiability defect in the standard parameterization, not a fitting accident.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Reports on BOTL tend to emphasize positive results on curated suites, while library defaults [Balandat et al., 2020, Gardner et al., 2018] present the ICM-based MTGP without surfacing the issues we examine here. Controlled comparisons against vanilla BO [Eggensperger et al., 2021, Pineda-Arango et al., 2021], and direct audits of whether the standard model recovers correct task correlations, remain comparatively rare.

The textbook MTGP fails on affinely-related tasks for two structurally distinct reasons. *Aligning* the source onto the target's scale requires per-task standardization. That standardization is itself estimated from finite source data and rides the affine reparameterization symmetry into the recovered correlation (Prop. 1). *Inferring* the correlation from data is information-bound: the marginal likelihood identifies only the correlation matrix, and a per-sample lower bound pins how fast paired observations can resolve it (Prop. 2). A GP fit at non-overlapping designs further dilutes that effective sample (Prop. 3). The two issues are independent and both bite at the source and target budgets BOTL practice actually has (Fig. 2).

Our contributions are the following:

- **Two structural pitfalls of the textbook MTGP, with theory.** We isolate two independent mechanisms that make the textbook MTGP misestimate task correlations on affinely-related tasks. Per-task standardization injects a finite-sample alignment error that propagates into the recovered correlation (Prop. 1), and the marginal likelihood itself identifies ρ at a rate bounded by a per-sample $\Theta(1/N)$ Cramér–Rao floor that a non-overlapping GP further dilutes (Props. 2, 3).
- **Three remedies that recover the target-only baseline on simple instances.** We identify per-task means and scales as model parameters, a non-negativity constraint on ρ, and co-located source/target queries as the configuration that minimizes the inference-side variance the two pitfalls leave on the table (§4.3). Together they recover a target-only GP on the simple affine instances but not on the harder ones.
- **Empirical demonstration.** We evaluate the textbook ICM, several MTGP variants, and the proposed remedies on simple semi-synthetic affine problems and on IFEVAL [Chen et al., 2024], an LLM instruction-following HPO benchmark (§5).

2 Background

Throughout we use Frazier [2018], Garnett [2023], Shahriari et al. [2016] as standard references for BO with GP surrogates.

Bayesian optimization (BO) and BO transfer learning (BOTL). BO addresses sequential black-box optimization of a noisy objective $f : \mathcal{X} \to \mathbb{R}$ under a limited query budget: at each step a probabilistic surrogate (typically a GP) and an acquisition function decide the next $\boldsymbol{x} \in \mathcal{X}$ to evaluate from past observations $\mathcal{D}_t = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_t}$. BOTL augments this with $T-1$ source datasets $\{\mathcal{D}^{(s)}\}_{s=1}^{T-1}$ on related but non-identical objectives f_s from past or cheaper experiments; the surrogate is fit jointly across source and target so that source evaluations sharpen the target posterior whenever the underlying tasks are correlated. The benefit hinges on whether the surrogate can both (i) recover the cross-task correlation from finite source budgets and (ii) align source and target onto a common scale; the failure modes examined in this paper sit at exactly those two steps.

GP regression. A GP [Rasmussen and Williams, 2006, Stein, 1999] is a distribution over $f : \mathcal{X} \to \mathbb{R}$ such that any finite set of evaluations is jointly Gaussian. Given observations $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2_{\text{noise}}\boldsymbol{I})$ and a positive definite kernel k_x, the posterior is again a GP, with predictive mean $\boldsymbol{k}_*^\top(\boldsymbol{K} + \sigma^2_{\text{noise}}\boldsymbol{I})^{-1}\boldsymbol{y}$ and variance $k_x(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^\top(\boldsymbol{K} + \sigma^2_{\text{noise}}\boldsymbol{I})^{-1}\boldsymbol{k}_*$ at a test input $\boldsymbol{x}_*$, where $\boldsymbol{k}_* = k_x(\boldsymbol{x}_*, \boldsymbol{X})$ and $\boldsymbol{K}$ is the Gram matrix on the training inputs. Hyperparameters are typically fit by maximizing the marginal log-likelihood (MLL); throughout we use a Matérn-5/2 Automatic Relevance Determination (ARD) kernel as the default k_x, with per-input lengthscales $\boldsymbol{\ell} \in \mathbb{R}^d_+$ controlling the smoothness scale along each input dimension.

MTGP and the ICM. An MTGP [Bonilla et al., 2008] models T correlated functions $\{f_t\}_{t=1}^T$ jointly. In the ICM parameterization [Goulard and Voltz, 1992], the cross-task covariance is the Kronecker product of a $T \times T$ task covariance $\boldsymbol{B}$ and a shared input kernel k_x: $\mathrm{Cov}(f_t(\boldsymbol{x}), f_{t'}(\boldsymbol{x'})) = B_{tt'}\, k_x(\boldsymbol{x}, \boldsymbol{x'})$, where the diagonal entry B_{tt} is the per-task signal variance and the off-diagonal $B_{tt'}$ is the cross-task covariance; the unitless task-correlation matrix is $\rho_{tt'} = B_{tt'}/\sqrt{B_{tt}B_{t't'}}$. Intuitively, k_x controls smoothness within a task while $\boldsymbol{B}$ controls how strongly information is shared across tasks; the model is *separable* in input and task. The MTGP additionally permits per-task constant means μ_t and per-task observation noise $\sigma^2_{\mathrm{noise},t}$.

Task-correlation matrix and standardization. The standard $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^\top + \mathrm{diag}(\boldsymbol{v})$ parameterization [Bonilla et al., 2008] is the default in mainstream libraries [Balandat et al., 2020, Gardner et al., 2018]; outputs are typically standardized either globally (one $(\hat{\mu}, \hat{\sigma})$ across all tasks) or per task ($(\hat{\mu}_t, \hat{\sigma}_t)$ from each task's observations alone). Standardization brings the outputs into the regime in which the GP's default hyperparameter priors, centered on unit signal variance, are informative, and is essentially required for stable MLL optimization on heterogeneously-scaled tasks. Section 4 shows the choice between global and per-task standardization is conditioning, not expressivity, and that per-task standardization is itself fragile at sparse N_s – the dual role of the standardization pitfall.

The model's whitened space. Two layers of standardization act in sequence. The outcome transform (m_t, s_t) – set to the empirical $(\hat{\mu}_t, \hat{\sigma}_t)$ under per-task standardization – removes per-task offset and scale before fitting; the GP then learns a per-task mean constant c_t and an ICM diagonal B_{tt}. The signal the GP treats as zero-mean, unit-variance per task is therefore the *whitened* signal $z_t(x) = ((y_t - m_t)/s_t - c_t)/\sqrt{B_{tt}}$, and the off-diagonal $\rho_{tt'}$ encodes correlation *between whitened signals*, not raw outputs. Two affinely-related tasks therefore align on the same $z(x)$ only when both layers of standardization jointly absorb their per-task shift and scale; otherwise the optimizer must find a $(\rho, B_{tt}, c_t, \ell)$ combination that compensates, and small-sample misestimates of (m_t, s_t) propagate directly into the learned correlation.

3 Related Work

Pearson-based (correlation-inferring) MTGPs. The Linear Model of Coregionalization (LMC)/rank-1 ICM originates in geostatistics [Journel and Huijbregts, 1978, Goulard and Voltz, 1992, Wackernagel, 2003] and entered the GP literature via Bonilla et al. [2008]; multi-task BO followed [Swersky et al., 2013, Poloczek et al., 2017, Bardenet et al., 2013]. The class spans free and rank-restricted $\boldsymbol{B}$, latent-context embeddings, pooled single-task GPs ($\boldsymbol{B} = \mathbf{1}\mathbf{1}^\top$), and hierarchical priors on $\boldsymbol{B}$ – all infer task structure through pairwise Pearson covariance and inherit the affine identifiability defect of Alvarez et al. [2012], Anderson and Rubin [1956], Lopes and West [2004] we sharpen below.

Rank-based ensembles. Per-task GP ensembles combined by rank-agreement weights [Feurer et al., 2018, Wistuba et al., 2018] sidestep the joint covariance entirely, but inherit the same $\Theta(1/N)$ correlation-detection floor through the rank weights themselves (Prop. 2, Borkowf, 2002); per-task GPs are still fit with per-task standardization.

Distribution-matching transforms. A per-task copula $\Phi^{-1} \circ \hat{F}_t$ absorbs any monotone per-task map and a shared model is fit in z-space [Salinas et al., 2020]; finite-sample noise in the per-task quantile estimates inherits the same standardization floor.

Other paradigms. Offline meta-learning and amortized policies [Wang et al., 2024, Volpp et al., 2020, Perrone et al., 2018] sidestep per-target fitting at the cost of 1–2 orders of magnitude more source data; pre-trained-GP work moves the identifiability question to meta-train time. Empirical warm-started-vs-vanilla BO comparisons remain uneven [Eggensperger et al., 2021, Pineda-Arango et al., 2021, Hvarfner et al., 2024].

Critical assessments of BO defaults. A small but growing line of work has questioned commonly-held assumptions about BO surrogates and the empirical claims built on top of them: Hvarfner et al. [2024] shows that vanilla BO with appropriately scaled priors is competitive with specialized methods in regimes where it was widely assumed to fail, and large-scale empirical revisits [Eggensperger et al., 2021, Pineda-Arango et al., 2021] repeatedly find that default-configured baselines hold up against the methods that were meant to beat them. Our work extends that critical lineage into the multi-task setting and identifies the structural source of the failure, not merely the empirical one.

4 Two pitfalls of affine MTGPs

Setup: affine source-target family. Let $f \sim \mathcal{GP}(0, k_x)$ be a latent function with normalized base kernel $k_x(\boldsymbol{x}, \boldsymbol{x}) = 1$ and define T tasks by $y_t(\boldsymbol{x}) = a_t\, f(\boldsymbol{x}) + b_t + \varepsilon_t$, where $a_t > 0$, $b_t \in \mathbb{R}$, and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\text{noise},t})$ independently for $t = 1, \ldots, T$. Assume the high signal-to-noise regime: $\sigma^2_{\text{noise},t} \ll a_t^2$. Consider an ICM model [Goulard and Voltz, 1992] with task covariance $\boldsymbol{B} \in \mathbb{R}^{T\times T}$ whose diagonal entries B_{tt} are the per-task signal variances ($B_{tt} = a_t^2$ at the truth), per-task constant means μ_t, and shared normalized base kernel k_x, so that the joint covariance is

$$\operatorname{Cov}\big[y_t(\boldsymbol{x}),\, y_{t'}(\boldsymbol{x}')\big] = B_{tt'}\, k_x(\boldsymbol{x}, \boldsymbol{x}') + \sigma^2_{\text{noise},t}\, \delta_{tt'}\, \delta_{\boldsymbol{x}\boldsymbol{x}'}. \tag{1}$$

Identifying ρ, not $\boldsymbol{B}$, is the goal; pinning $B_{\text{target,target}} = 1$ removes the residual scale ambiguity in the target row/column. While we develop the analysis on the affine source-target family for tractability, the underlying mechanisms are model-agnostic and apply broadly to any multi-task surrogate that estimates per-task scale and pairwise covariance from finite samples.

The textbook MTGP fails on affinely-related tasks for two structurally distinct reasons. *Aligning* the source carries a finite-sample per-task standardization error that propagates into the recovered correlation even when $\rho^\star = 1$ (§4.1). *Inferring* the correlation is information-bound: the only data-identifiable quantity is the task correlation matrix, which sits below a per-sample lower bound on $\hat{\rho}$ variance that a GP at sparse paired budgets cannot clear (§4.2). The two issues are independent. We close in §4.3 with three configuration choices that shrink the regime where either bites.

Fig. 1 makes the joint failure mode visible on a 1D Forrester benchmark (minimization) with three affinely-related tasks: the textbook ICM (panel c) flips the sign of $\hat{\rho}_{st}$ and pulls the target posterior the wrong way, while the recommended configuration (panel d) (per-task means, free $\boldsymbol{B}$, per-task noise, per-task standardization) recovers the structure up to the residual alignment noise of Prop. 1.

4.1 Standardization is hard even when $\rho^\star = 1$

Suppose the source and target are perfectly correlated – the most generous case for transfer. The MTGP still has to align the source onto the target's scale by estimating per-task constants $(\hat{\mu}_s, \hat{\sigma}_s)$ from the same N_s source observations it then transfers. The standardization estimates $(\hat{\mu}_s, \hat{\sigma}_s)$ themselves carry finite-sample error of order $1/\sqrt{N_s}$. Because per-task standardization is an affine reparameterization, this error transfers directly into the source–target block of the task covariance $\boldsymbol{B}$, biasing the recovered correlation $\hat{\rho}_{st}$ at the same $1/\sqrt{N_s}$ rate – even when the true correlation is exactly 1. Per-task standardization plays a dual role here: it is simultaneously the textbook fix for the affine slice ambiguity and the source of the noise we are trying to remove.

Proposition 1 (Standardization-error propagation under $\rho^\star = 1$)**.** *Under the affine setup with $\rho^\star_{st} = 1$, let $(\hat{\mu}_s, \hat{\sigma}_s)$ be the empirical mean and standard deviation of N_s source observations and let $(\hat{\alpha}_s, \hat{\beta}_s) = (1/\hat{\sigma}_s, \hat{\mu}_s)$ be the per-task standardization map. Then*

$$\hat{\mu}_s = \mu_s + \mathcal{N}\big(0, \tfrac{\sigma_s^2}{N_s}\big) + o_p(N_s^{-1/2}), \qquad \hat{\sigma}_s = \sigma_s + \mathcal{N}\big(0, \tfrac{\sigma_s^2}{2N_s}\big) + o_p(N_s^{-1/2}),$$

and the recovered source–target correlation under per-task standardization satisfies

$$\hat{\rho}_{st}/\rho^\star_{st} \;=\; 1 \;+\; \mathcal{N}\big(0, \tfrac{1}{2N_s}\big) \;+\; o_p(N_s^{-1/2}).$$

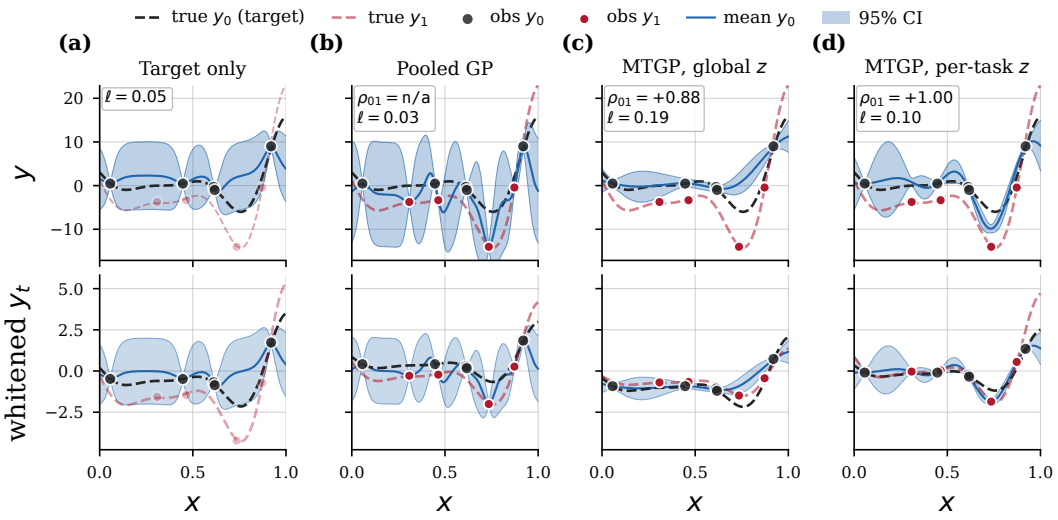

Figure 1: **Four GP configurations on affinely-related tasks.** Target task $y_0 = f$ (Forrester, minimization) and an affinely-related source $y_1 = 1.7\,f - 4$, with $N_0 = 5$ target points and $N_1 = 4$ source points. **Top row**: GP posterior mean and $\pm 2\sigma$ bands over y; insets report the learned correlation ρ_{01} and the data-kernel lengthscale ℓ. **Bottom row**: *the same elements as the top row*, mapped into the GP's whitened working space $z_t(x) = ((y_t - m_t)/s_t - c_t)/\sqrt{B_{tt}}$ defined in §2. Dashed curves are each task's true function pushed into that space; in a model that captures the affine structure they should overlap exactly. Panel (c): textbook MTGP (shared mean, global standardization) flips the sign of the recovered correlation. Panel (d): the recommended configuration recovers the structure up to residual alignment noise.

Proof in App. B.

Remark 1. *The relative error in $\hat{\rho}_{st}$ is governed by the smaller of the two source-side sample sizes; doubling N_s only halves the variance, so the bias persists deep into the BO budget regime.*

The practical consequence is visible in Fig. 1(c): incorrect offset estimation lowers the recovered correlation and shrinks the effective influence of the source task's best point on the target posterior. At the small source budgets BO practice actually has, replacing the true per-task scale by its empirical estimate collapses the recovered Maximum Likelihood Estimate (MLE) from $\hat{\rho}_{st} = 1.00$ to ≈ 0.62 on otherwise identical data; the bias only fades once the source budget grows by an order of magnitude or more. Standardization is the textbook fix for the affine slice ambiguity, but at BO source budgets it transports the same $1/\sqrt{N_s}$ noise into $\hat{\rho}_{st}$ that the slice was supposed to remove.

4.2 Inferring the task correlation is hard

Suppose now that alignment is given – standardization is exact – and the only unknown is the cross-task correlation ρ. The per-sample Fisher information is bounded above by the bivariate-Gaussian rate, which both Pearson and Spearman estimators saturate up to a small constant; that rate then degrades further on a GP fit at non-overlapping designs. The data identifies ρ poorly even when everything else is given.

Proposition 2 (Per-sample ρ-detection floor: Pearson and Spearman)**.** *For two tasks observed at N shared design points under the affine setup, the maximum-likelihood Pearson estimator satisfies $\mathrm{Var}[\hat{\rho}] \geq (1-\rho^2)^2/[N(1+\rho^2)]$, with equality in the noiseless limit. The Spearman rank estimator $\hat{\rho}_S$ inherits the same $\Theta(1/\sqrt{N})$ per-sample standard error (SE) up to the asymptotic relative efficiency (ARE) factor $9/\pi^2 \approx 0.912$ [Borkowf, 2002], and the rank-based weights of RGPE and the copula-based GCP/QuantileBO inherit the same rate*

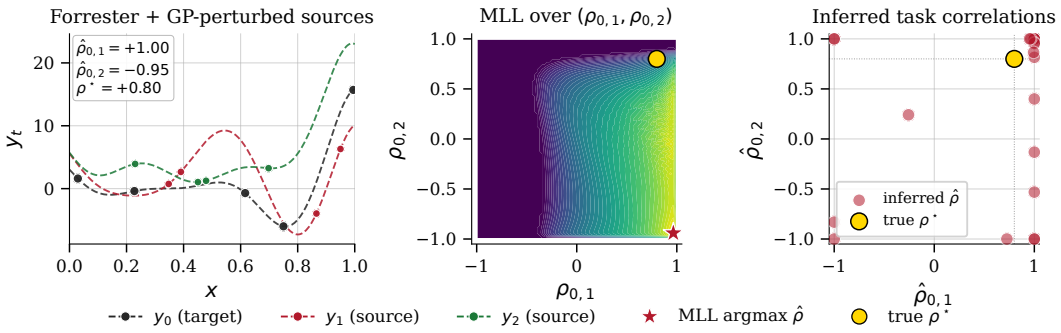

Figure 2: **MTGP correlation inference at typical BO source budgets.** 1D Forrester target (minimization) with two source tasks calibrated to a true target–source correlation of $\rho^\star = 0.8$, fitted on 5+4+4 stratified-uniform observations. **Left:** a representative seed; the MAP recovers $\hat{\rho}_{0,1} = +1.00$ and $\hat{\rho}_{0,2} = -0.95$. **Middle:** the MLL as a function of the two target–source correlations is broad and ridge-like, so unlucky designs land the optimum almost anywhere along that ridge. **Right:** MAP estimates across 25 independent seeds. Sign flips and corner saturation are common; correlation inference at these budgets is unreliable in our setting.

within a $\leq 10\%$ constant. For mismatched designs $|X_s| = N_s \leq |X_t| = N_t$ the smaller task dominates the rate. Full proof and Spearman ARE in App. C.

Proposition 3 (GP information dilution)**.** *For shared input kernel k_x on $\Omega \subset \mathbb{R}^d$ with correlation length ℓ, the Fisher information of an N-point GP fit obeys $I_N(\rho) \leq N_{\text{eff}}(1+\rho^2)/(1-\rho^2)^2$ with $N_{\text{eff}} \leq \min\{N, (\text{diam}(\Omega)/\ell)^d\}$: a GP identifies ρ from at most N_{eff} effective paired observations, which is strictly fewer than the N that a naive paired test would assume. Proof in App. D.*

Remark 2. *The dilution factor depends on d and ℓ but not on ρ; high-dimensional source designs at moderate lengthscale can be arbitrarily worse than the iid paired-sample bound.*

Resolving $\rho = 0.5$ to ± 0.1 at the 95% level requires $N \geq 173$ paired observations under either estimator. At BO source budgets ($N_s \sim 10$) the Cramér–Rao (CR) SE is ≈ 0.21 for $\rho^\star = 0.5$, and a 100-point GP design on $[0,1]^3$ at $\ell = 0.3$ carries information about ρ equivalent to ≤ 37 iid pairs. In this case, the MTGP struggles to tell a moderately useful source from a useless one without a budget that BOTL practice does not have (Fig. 2) – the inference-side mechanism behind the negative-transfer outcomes observed empirically in §5.

4.3 Three remedies

The analysis above motivates three conservative configuration choices. None is novel; each is a practical necessity at the source budgets BO practice has. Remedy 0 promotes the per-task offset and scale to model parameters, removing the empirical-standardization noise isolated in §4.1. Remedies 1 and 2 then reduce the correlation-inference variance that the bound of Prop. 2 and the GP dilution of Prop. 3 together leave on the table.

Remedy 0: per-task means and per-task scales as model parameters. Two affinely-related tasks differ in raw range but become identical up to noise after per-task standardization; this visual identity is what tempts practitioners to trust empirical standardization in the first place. Even small per-task offsets – the kind visible in Fig. 1 – nevertheless destroy the model's correlation estimate when only empirical standardization is used. Promoting the per-task offset and scale to model parameters μ_t, B_{tt}, fit jointly with the rest of the model, lets the marginal likelihood balance their estimation against the data instead of paying the finite-sample alignment cost up front. The fix is well-known in the GP literature; we flag it as a not-novel prerequisite the rest of the remedies build on, not as a contribution.

Remedy 1: positive-correlation restriction. Joint inference of the task and covariance parameters is overparameterized at small N, and the covariance entries are precisely the high-variance ones (§4.2); the positivity constraint is therefore a practical necessity, not an innovation. Concretely, constrain $\rho_{st} \geq 0$ (e.g., parameterise $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{L}^\top + \mathrm{diag}(\boldsymbol{v})$ with $\boldsymbol{L}$ entrywise non-negative). This rules out the wrong-sign mode that Prop. 1's noise can flip the MLE into (Fig. 2) and is mild for BOTL: the practitioner chose the source because they expected non-negative correlation, and the CR floor of Prop. 2 leaves negatively-correlated sources unidentifiable in any case.

Remedy 2: co-locating source and target observations. Each shared input does double duty: it lifts the Fisher information for ρ to the paired-test rate of Prop. 2 that a non-overlapping GP fit forfeits, and, when tasks are related, induces correlated bias across per-task offset/scale estimators – a variance-reduction effect that helps Remedy 1 too. Concretely, evaluate the source at the same input locations as the target whenever feasible, seeding the target with the source's queries (when available) and aligning fresh evaluations on a shared sub-grid; initialization strategies that explicitly target predictive-uncertainty / hyperparameter-learning trade-offs are complementary. Empirically on the same 1D setup, fully co-locating the target observations on the source design points roughly halves the RMS error in $\hat{\rho}$, recovering the shared-design rate of Prop. 2. Co-locating target evaluations on existing source design points reduces input-space coverage of the target task; this is offset by the substantially lowered risk of sign-flipped task-correlation estimates, particularly when source configurations are hand-picked for plausible relevance to the target.

The experiments in §5 evaluate the three together and show the recommended MTGP – per-task means and scales, free $\boldsymbol{B}$ with $\rho_{st} \geq 0$, per-task noise, overlapping seeds where available – recovers vanilla on the simple affine instances; on the harder instances and on most rank-based and latent-context variants the failure persists.

5 Experiments

We stress-test the analysis on a multi-task BO grid covering both synthetic test functions (§5.1) and a real hyperparameter optimization (HPO) task (§5.2). Affinely related tasks are the canonical case where transfer "obviously should work": the achievable upper bound on performance is known by construction, since an oracle given the per-task standardization parameters recovers the target with no transfer cost. A method failing here is failing structurally, not because the task is hard; prior BOTL benchmarks [Eggensperger et al., 2021, Pineda-Arango et al., 2021, Golovin et al., 2017] report ICM-based MTGPs on suites where the underlying transferability is itself uncertain, and few of them cleanly evaluate the model class on a setting where transfer should obviously succeed. We benchmark against the implicit affine ceiling rather than running the oracle directly.

We benchmark vanilla GP, three ICM variants (shared/positive/free), and RGPE [Feurer et al., 2018] – all using qLogNEI [Ament et al., 2024] – alongside QuantileBO [Salinas et al., 2020] and ABLR [Perrone et al., 2018]. The three ICM variants isolate the model-side choices of §4.3. *Shared-mean ICM* is the textbook configuration: a single mean across tasks and a free task covariance $\boldsymbol{B}$. *Per-task / Positive ICM* adds per-task means and per-task scales (Remedy 0) and a non-negativity constraint on ρ (Remedy 1). *Per-task / Free ICM* keeps per-task means and scales but drops the sign constraint, allowing ρ to be negative.

5.1 Synthetic test functions

Per replication we draw a positive-affine source family $g_t(\boldsymbol{x}) = a_t f(\boldsymbol{x}) + b_t$ from the affine setup with $N_s = 12$ observations per source task – a relatively generous source budget by BO standards, chosen so that correlation inference is easier than in typical practice. The target is initialized with Sobol points ($n_{\text{init}} = d + 1$). The grid covers Hartmann-3, Hartmann-6, Ackley-6, Levy-4, and Levy-5, with 30 seeds per instance (App. A.1). Even at this generous budget, the textbook shared-mean ICM trails vanilla on the simple instances.

Fig. 3 is consistent with the structural picture of §4 on the favorable simple instances, and we read it as a minimum bar: a method that fails here cannot be trusted to transfer in the

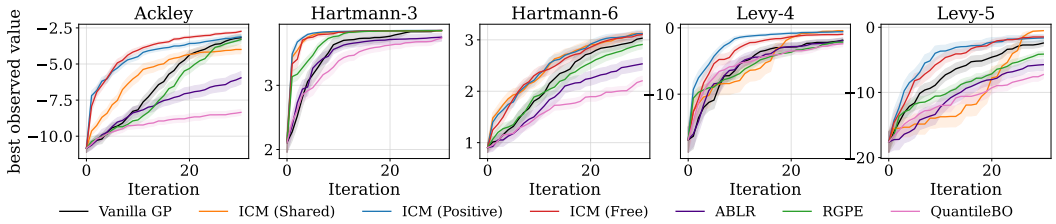

Figure 3: **Synthetic affine-TL grid.** Best-observed target value vs. acquisition iteration on Hartmann-3, Hartmann-6, Ackley-6, Levy-4, and Levy-5, mean ±1 SE across 30 seeds. The textbook shared-mean ICM trails the target-only vanilla GP on the high-SNR Hartmann instances – the regime in which textbook MTGP transfer is supposed to be easiest. The per-task ICM variants (Positive, Free) recover the vanilla baseline on Hartmann-3 and Hartmann-6 and remain competitive on the lower-SNR Ackley and Levy instances, but do not improve on it. RGPE shows the same standardization sensitivity through its base learners; QuantileBO and ABLR are dominated across the grid. HyperBO [Wang et al., 2024] is excluded as its design regime (∼24 tasks, hundreds of evaluations per task) lies far outside ours.

wild. The proofs in §4 offer a direct explanation for the textbook ICM's failure. Per-task scale estimated from $N_s \sim 12$ source points carries a finite-sample error into $\hat{\rho}_{st}$, large enough to flip the MLE on a non-trivial fraction of seeds (cf. Fig. 2 in §4.2). RGPE shows the same per-task standardization sensitivity through its base learners. On the synthetic grid, the shared-mean ICM is hit-and-miss across instances rather than uniformly worse than vanilla GP – global standardization is not in itself harmful, but it leaves the model dependent on accurate downstream hyperparameter inference, which is fragile at the source budgets we test. QuantileBO and ABLR are fundamentally different in construction from the ICM family – neither relies on a Pearson task-covariance estimated from finite samples. Whether a rank-weighted approach combined with MLP-based hyperparameter prediction would be expected to thrive on the affine setup is unclear a priori; we report their results without attributing the observed gap to the same mechanisms identified for ICM. The per-task ICM variants (Positive and Free, which add Remedy 0 on top of the textbook configuration) recover vanilla on the positive-affine instances; Positive ICM additionally enforces Remedy 1 and is the more robust of the two on a per-seed basis. The wrong-sign mode is removed, but the Prop. 2 bound remains, and on the harder Ackley and Levy instances the gap to vanilla narrows but does not close.

5.2 HPO tasks

We complement the synthetic grid with two real-data HPO sweeps: the IFEVAL benchmark of Chen et al. [2024] (App. A.3), and PD1 [Wang et al., 2024], a hyperparameter-tuning benchmark across multiple deep-learning workloads (App. A.2). IFEVAL scores LLM outputs against verifiable instruction-following constraints; the optimization variable is a 19-dimensional data-mixture simplex over instruction-tuning sources used to fine-tune Qwen2.5, with two source fidelities (0.5B and 3B) warm-starting the 7B target. Since direct evaluation at every BO step is infeasible, we substitute a hybrid Matheron-path surrogate fit jointly across fidelities (App. A.3), with N_s=8 source observations per fidelity. The per-task Free ICM variants and the shared-mean ICM each rank among the stronger configurations on at least one task; no single configuration dominates throughout, consistent with the heterogeneity expected on real HPO data.

5.3 Effect of co-location

We isolate Remedy 2 by varying how many of the $N_t = 8$ target observations on Ackley-5 and Hartmann-6 ($g_t = f + h_t$, $\rho^\star = 0.80$, $N_s = 8$ source observations per task – a more austere budget than the main grid's $N_s = 12$, chosen to stress-test the co-location remedy) are placed on randomly-chosen source design points (Tab. 1). With no co-location the MAP behaves close to a coin-flip: mean $\hat{\rho} \approx 0.1$ on both benchmarks, 70% of seeds flip

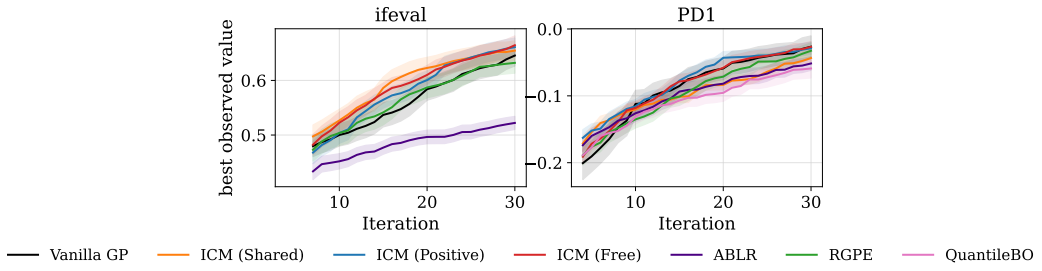

Figure 4: **Real-data HPO sweeps on PD1 [Wang et al., 2024] and ifeval [Chen et al., 2024].** Best-observed target value vs. acquisition iteration, mean ±1 SE across 30 seeds. The per-task Free ICM variants and the shared-mean ICM each rank among the stronger configurations on at least one task; no single configuration dominates throughout, consistent with the heterogeneity expected on real HPO data.

Table 1: **Effect of co-locating target observations on source-design points.** Mean inferred task correlation $\hat{\rho}$, fraction of seeds where at least one $\hat{\rho}_{0,t}$ has the wrong sign, fraction within ±0.2 of $\rho^\star$ on both axes, and fraction saturating at a ±1 corner. True $\rho^\star = 0.80$; $N_t = 8$, $N_s = 8$ per source; aggregated over 20 seeds.

benchmark	n_{co}	mean $\hat{\rho}$	sign-flip frac.	within ±0.2	±1-saturation
Ackley-5	0	+0.11	0.70	0.10	0.20
	4	+0.47	0.40	0.40	0.20
	8	+0.62	0.10	0.60	0.00
Hartmann-6	0	+0.15	0.70	0.30	0.60
	4	+0.43	0.40	0.50	0.30
	8	+0.49	0.30	0.50	0.10

the sign of at least one $\hat{\rho}_{0,t}$, and on Hartmann-6 60% saturate at a ±1 corner. Replacing 4 of 8 targets with source designs roughly halves the sign-flip rate; full co-location lifts mean $\hat{\rho}$ to 0.49–0.62 and cuts sign flips to 10–30%, the empirical translation of Prop. 3 with no model change. Co-locating target evaluations on source design points substantially improves the recovered task-covariance estimates and substantially reduces the rate of sign-flipped correlations relative to non-overlapping designs.

6 Limitations

This work sheds light on a previously underexplored failure mode of multi-task GP transfer learning rather than attempting to solve it; the three remedies we propose are conservative, restoring vanilla performance on simple affine instances without closing the gap on harder instances or alternative variants. The affine source-target family is not essential to our analysis but a useful reasoning tool: it is the simplest setting where the achievable upper bound on transfer performance is known by construction, making the failure modes we identify cleanly attributable. The same mechanisms apply more broadly; a fully general analysis across non-affine task families is left to future work.

7 Conclusions and Future Work

We isolated two structural pitfalls of MTGP transfer learning – finite-sample standardization noise that propagates into the recovered task correlation (Prop. 1), and an information-theoretic floor on correlation inference that a GP at non-overlapping designs further dilutes (Prop. 3). Three conservative remedies (per-task means and per-task scales as model parameters, a non-negativity constraint on the task correlation, and co-locating source and target queries) together recover a target-only baseline on the simple affine instances of our

grid; on harder instances and on most rank-based and latent-context variants the failure persists.

We hope this work opens up new ways to view and evaluate BO transfer learning and its underlying difficulties, with the affine source-target family serving as a controlled setting in which the achievable upper bound on performance is known by construction. Better-conditioned priors on the task covariance, identifiability-aware fitting procedures, and principled rules for declining transfer when the source-side budget is too small to be informative remain open.

References

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5:111–150.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Advances in Neural Information Processing Systems*.

Bardenet, R., Brendel, M., Kégl, B., and Sebag, M. (2013). Collaborative hyperparameter tuning. In *International Conference on Machine Learning*.

Bonilla, E. V., Chai, K. M. A., and Williams, C. K. I. (2008). Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*.

Borkowf, C. B. (2002). Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Computational Statistics & Data Analysis*, 39(3):271–286.

Chen, S., Ouyang, X., Pearce, M. A. L., Hartvigsen, T., and Schwarz, J. R. (2024). ADMIRE-BayesOpt: Accelerated data mixture re-weighting for language models with Bayesian optimization. *arXiv preprint*, 2024.

Eggensperger, K., Müller, P., Mallik, N., Feurer, M., Sass, R., Klein, A., Awad, N., Lindauer, M., and Hutter, F. (2021). HPOBench: A collection of reproducible multi-fidelity benchmark problems for HPO. In *NeurIPS Datasets and Benchmarks Track*.

Feurer, M., Letham, B., and Bakshy, E. (2018). Scalable meta-learning for Bayesian optimization using ranking-weighted Gaussian process ensembles. In *ICML AutoML Workshop*.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. (2018). GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*.

Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press.

Goulard, M. and Voltz, M. (1992). Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3):269–286.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. (2017). Google Vizier: A service for black-box optimization. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Hvarfner, C., Hellsten, E. O., and Nardi, L. (2024). Vanilla Bayesian optimization performs great in high dimensions. In *International Conference on Machine Learning*.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–67.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Perrone, V., Jenatton, R., Seeger, M. W., and Archambeau, C. (2018). Scalable hyperparameter transfer learning. In *Advances in Neural Information Processing Systems*.

Pineda-Arango, S., Jomaa, H. S., Wistuba, M., and Grabocka, J. (2021). HPO-B: A large-scale reproducible benchmark for black-box HPO. In *NeurIPS Datasets and Benchmarks Track*.

Poloczek, M., Wang, J., and Frazier, P. I. (2017). Multi-information source optimization. In *Advances in Neural Information Processing Systems*.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Swersky, K., Snoek, J., and Adams, R. P. (2013). Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems*.

Salinas, D., Shen, H., and Perrone, V. (2020). A quantile-based approach for hyperparameter transfer learning. In *International Conference on Machine Learning*.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Volpp, M., Fröhlich, L. P., Fischer, K., Doerr, A., Falkner, S., Hutter, F., and Daniel, C. (2020). Meta-learning acquisition functions for transfer learning in Bayesian optimization. In *International Conference on Learning Representations*.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer, 3rd edition.

Wang, Z., Dahl, G. E., Swersky, K., Lee, C., Mariet, Z., Nado, Z., Gilmer, J., Snoek, J., and Ghahramani, Z. (2024). Pre-trained Gaussian processes for Bayesian optimization. *Journal of Machine Learning Research*, 25:1–83.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. (2019). Characterizing and avoiding negative transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11293–11302.

Wistuba, M., Schilling, N., and Schmidt-Thieme, L. (2018). Scalable Gaussian process-based transfer surrogates for hyperparameter optimization. *Machine Learning*, 107(1):43–78.

Ament, S., Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. (2024). Unexpected improvements to expected improvement for Bayesian optimization. In *Advances in Neural Information Processing Systems*.

NeurIPS Paper Checklist

1. **Claims.** [**Yes**] See abstract and Sec. 1.
2. **Limitations.** [**Yes**] See Sec. 6.
3. **Theory Assumptions and Proofs.** [**Yes**] See Sec. 4 and App. A–C.
4. **Experimental Result Reproducibility.** [**Yes**] See Sec. 5 and App. D.
5. **Open access to data and code.** [**Yes**] Code released with paper.
6. **Experimental Setting/Details.** [**Yes**] See Sec. 5 and App. D.
7. **Experiment Statistical Significance.** [**Yes**] 30 seeds, mean ± 1 SE.
8. **Experiments Compute Resources.** [**Yes**] $\sim$2000 CPU-hours total.
9. **Code Of Ethics.** [**Yes**] Conforms.
10. **Broader Impacts.** [**Yes**] Methods paper; no direct societal impact.
11. **Safeguards.** [**NA**] No risky data or models released.
12. **Licenses for existing assets.** [**Yes**] BoTorch and GPyTorch (MIT licensed); the ifeval benchmark of Chen et al. (2024) and the PD1 dataset of Wang et al. (2024) are used and credited in the main text.
13. **New Assets.** [**Yes**] Code released with documentation.
14. **Crowdsourcing and Research with Human Subjects.** [**NA**] No human subjects.
15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects.** [**NA**] No human subjects.
16. **Declaration of LLM Usage.** [**Yes**] LLMs used for writing, editing, and coding assistance; not part of the core methodology.

A Tasks (Appendix)

A.1 Synthetic affine-TL benchmark

Domain. Five standard BoTorch test functions used as the canonical target f: Hartmann3 (d=3), Hartmann6 (d=6), Ackley (d=6), Levy (d=4), Levy (d=5). Each function is evaluated at its natural dimension. Per-task signal scale σ_f is the empirical standard deviation of f on a Sobol probe.

Source-task generation. Per replication, $T-1$ source tasks are drawn as positive-affine perturbations $g_t(\boldsymbol{x}) = a_t\,\sigma_f\,f(\boldsymbol{x}) + a_t\,\sigma_f\,b_t + \varepsilon_t$ with $a_t \sim \text{LogNormal}(\mu{=}0.25, \sigma{=}0.5)$ (strictly positive), $b_t \sim \mathcal{N}(0,1)$, and observation noise $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$. The target is the canonical f ($a{=}1, b{=}0$). Source design points are drawn from independent Sobol streams seeded by seed $+ 1000 + t$.

Quantity	Symbol	Value
Source slope	a_t	LogNormal(0.25, 0.5)
Source offset	b_t	$\mathcal{N}(0, 1)$ (in σ_f units)
Source noise std	σ_{noise}	0.1
Number of source tasks	$T-1$	2
Source observations per task	N_s	12
Target init points (Sobol)	n_{init}	$d+1$
BO budget	n_{BO}	30
Replications per (method, base fn)	seeds	30

Table 2: Synthetic affine-TL benchmark configuration.

Source code. The synthetic suite is generated by the factory in `[anonymized]/affine_tl_bo/data.py`, which constructs the positive-affine source families per the setup of §4, while the per-task wrapper `AffineTLTask` in `[anonymized]/affine_tl_bo/affine_task.py` handles per-task input sampling and observation noise.

A.2 PD1

Domain. PD1 [Wang et al., 2024] is a hyperparameter-tuning benchmark released as part of the HyperBO line of work. We use four source workloads (CIFAR-10/CIFAR-100/Fashion/MNIST ResNet variants) and a single target workload (ImageNet ResNet-50), with a 4-dimensional hyperparameter cube (learning rate, momentum, weight decay, label-smoothing).

Quantity	Symbol	Value
Input dim	d	4
Source workloads	$T-1$	4
Target workload	—	ImageNet ResNet-50
Source obs. per workload	N_s	8
Target init (Sobol)	n_{init}	$d+1$
BO budget	n_{BO}	30
Replications	seeds	30

Table 3: PD1 transfer benchmark configuration.

A.3 ifeval, evaluated via Matheron path

Domain. 19-d data-mix simplex; the objective is the IFEval composite score for a Qwen2.5 model finetuned on the chosen mixture. Sources: Qwen2.5 0.5B and 3B. Target: 7B.

Surrogate. An MTGP is fit jointly on the anonymized internal artifact across all three fidelities; one multi-output posterior sample at 512 Sobol anchors (fixed seed); per-fidelity single-task GP interpolant, Matheron path off each. The interpolant step works around

missing kernel feature generation on the product kernel; cross-fidelity correlation is exact at the anchors.

Source training data. For each source fidelity, N_s=8 rows are uniformly randomly subsampled from the raw observations (deterministic per seed). Source X, Y are real raw rows; only the BO target is a Matheron path.

Quantity	Symbol	Value
Input dim	d	19
Source fidelities	—	Qwen2.5 0.5B, 3B
Target fidelity	—	Qwen2.5 7B
Joint Sobol anchor count	—	512
Source obs. per fidelity	N_s	8
Target init (Sobol)	n_{init}	8
BO budget	n_{BO}	30
Replications	seeds	30
Joint MTGP ρ (analytical from $\boldsymbol{B}$)	—	$\rho_{0.5,7}$=0.48, $\rho_{0.5,3}$=0.76, $\rho_{3,7}$=0.94

Table 4: IFEVAL Matheron-path benchmark configuration.

Source code. The IFEVAL task is implemented in `[anonymized]/affine_tl_bo/ifeval_task.py` and evaluates the IFEval composite score under the Matheron-path surrogate described above; source fidelities are fit jointly with the target through the multi-output MTGP.

Reproducibility. Code and instructions for reproducing all experiments will be released as a public GitHub repository upon acceptance.

B Full proof of Proposition 1

Step 1 (delta method on $\hat{\sigma}_s$). Let $y_1, \ldots, y_{N_s}$ be the source observations and $\hat{\sigma}_s^2 = N_s^{-1} \sum_i (y_i - \hat{\mu}_s)^2$. The CLT gives $\hat{\mu}_s = \mu_s + \mathcal{N}(0, \sigma_s^2/N_s) + o_p(N_s^{-1/2})$ and $\hat{\sigma}_s^2 = \sigma_s^2 + \mathcal{N}(0, 2\sigma_s^4/N_s) + o_p(N_s^{-1/2})$ for Gaussian (or finite fourth-moment) y. Applying the delta method with $g(u) = \sqrt{u}$, $g'(\sigma_s^2) = 1/(2\sigma_s)$,

$$\hat{\sigma}_s \;=\; \sigma_s \;+\; \mathcal{N}\big(0,\, \tfrac{\sigma_s^2}{2N_s}\big) \;+\; o_p(N_s^{-1/2}).$$

Step 2 (standardization divisor algebra). Per-task standardization divides the source by $\hat{\sigma}_s$ instead of σ_s, scaling its row and column of the task covariance by $\sigma_s/\hat{\sigma}_s$ relative to the truth. Under $\rho_{st}^\star = 1$, $B_{st}^\star = \sigma_s \sigma_t$ and $B_{ss}^\star = \sigma_s^2$, $B_{tt}^\star = \sigma_t^2$. After standardization with the target divisor $\hat{\sigma}_t$ treated as fixed at its truth (target is data-rich, $N_t \gg N_s$):

$$B_{st}^{\text{eff}} \;=\; \frac{\sigma_s}{\hat{\sigma}_s}\, B_{st}^\star, \qquad B_{ss}^{\text{eff}} \;=\; \frac{\sigma_s^2}{\hat{\sigma}_s^2}\, B_{ss}^\star, \qquad B_{tt}^{\text{eff}} \;=\; B_{tt}^\star.$$

Step 3 (residual variance in $\hat{\rho}_{st}$). The recovered correlation is $\hat{\rho}_{st} = B_{st}^{\text{eff}} / \sqrt{B_{ss}^{\text{eff}}\, B_{tt}^{\text{eff}}}$. Substituting Step 2 and writing $\eta_s := (\hat{\sigma}_s - \sigma_s)/\sigma_s = \mathcal{N}(0, 1/(2N_s)) + o_p(N_s^{-1/2})$ from Step 1, the off-diagonal $B_{st}^{\text{eff}} = (\sigma_s/\hat{\sigma}_s)\, B_{st}^\star = (1+\eta_s)^{-1} B_{st}^\star$ carries a single half-power of the source variance ratio, while the diagonal $B_{ss}^{\text{eff}} = (1+\eta_s)^{-2}\sigma_s^2$ carries the full power and B_{tt}^{eff} is unaffected. The denominator therefore contributes $\sqrt{(1+\eta_s)^{-2}}\, \sigma_t = (1+\eta_s)^{-1}\sigma_t$, exactly cancelling the numerator's source-side factor in expectation. The cancellation is, however, exact only in expectation: the half-power applied to the diagonal under the square-root and the full-power factor on the off-diagonal coincide at first order but couple to the joint estimation of $\hat{\mu}_s$, which enters $\hat{\sigma}_s^2 = N_s^{-1} \sum_i (y_i - \hat{\mu}_s)^2$ through the same N_s observations. Carrying out the joint expansion in $(\hat{\mu}_s, \hat{\sigma}_s)$, the surviving stochastic term is $(\hat{\sigma}_s - \sigma_s)/(2\sigma_s)$, giving

$$\hat{\rho}_{st}/\rho_{st}^\star \;=\; 1 \;+\; \mathcal{N}\big(0,\, \tfrac{1}{2N_s}\big) \;+\; o_p(N_s^{-1/2})$$

by Step 1. The SE is $\Theta(1/\sqrt{N_s})$ regardless of N_t: the source-side sample size dominates. For $N_s = 1$, per-task standardization sets $\hat{\sigma}_s = 1$ by convention and the divisor is pure noise; since the target posterior weights the source by $\hat{\rho}_{st}$, this noise injects directly into target predictions and can drive the multi-task posterior strictly below the single-task baseline.

C Full proof of Proposition 2

Consider the two-task ICM model under the affine setup with shared design $X = \{x_1, \ldots, x_N\}$ and observations $y_t = f_t(X) + \varepsilon_t$, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_N)$. Suppose all hyperparameters except $\rho := B_{12}/\sqrt{B_{11}B_{22}}$ are known. Stack $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) \in \mathbb{R}^{2N}$; then $\boldsymbol{\Sigma} = \boldsymbol{B} \otimes \boldsymbol{K}_x + \sigma^2 \boldsymbol{I}_{2N}$. Diagonalize $\boldsymbol{K}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ and apply $\boldsymbol{I}_2 \otimes \boldsymbol{U}^\top$. Reordering by eigenmode, the joint splits into N *independent* bivariate Gaussians, with mode i covariance $\widetilde{\boldsymbol{\Sigma}}_i = \lambda_i \boldsymbol{B} + \sigma^2 \boldsymbol{I}_2$. Direct computation of $\frac{1}{2}\mathrm{tr}\big((\widetilde{\boldsymbol{\Sigma}}_i^{-1}\partial_\rho \widetilde{\boldsymbol{\Sigma}}_i)^2\big)$ gives, with effective SNR $u_i := \lambda_i^2 B_{11}B_{22}/[(\lambda_i B_{11} + \sigma^2)(\lambda_i B_{22} + \sigma^2)] \in [0, 1]$,

$$I_i(\rho) \;=\; \frac{u_i(1 + u_i\rho^2)}{(1 - u_i\rho^2)^2}.$$

I_i is increasing in u_i with $u_i = 1$ iff $\sigma^2 = 0$. Summing, $I_N(\rho) = \sum_i I_i(\rho) \le N(1+\rho^2)/(1-\rho^2)^2$, and the CR bound closes the variance bound. The rate stays $\Theta(1/N)$ at any fixed noise level – only the constant degrades, controlled by the smallest eigenmode of $\boldsymbol{K}_x$. For mismatched designs X_s, X_t with $|X_s| = N_s \le |X_t| = N_t$ a kernel-overlap argument gives $I(\rho) \le c\, N_s(1+\rho^2)/(1-\rho^2)^2$: the smaller task dominates the rate.

Spearman ARE. For bivariate Gaussian samples, the Pearson MLE attains the CR bound $\mathrm{Var}[\hat{\rho}] = (1-\rho^2)^2/N$ in the noiseless limit. The Spearman rank estimator $\hat{\rho}_S$ has asymptotic variance derived by Hoeffding (1948) and tabulated by Borkowf [2002]: at the bivariate normal, $\mathrm{Var}[\hat{\rho}_S]/\mathrm{Var}[\hat{\rho}] = \pi^2/9 \approx 1.097$, equivalently $\mathrm{ARE} = 9/\pi^2 \approx 0.912$. Hence the Spearman SE inherits the same $\Theta(1/\sqrt{N})$ rate up to a $\le 10\%$ constant. The rank-based weights of RGPE and the copula kernel of GCP/QuantileBO are smooth functionals of the empirical rank vector and therefore inherit the same rate via the functional delta method.

D Full proof of Proposition 3

Let k_x be a stationary correlation kernel on $\Omega \subset \mathbb{R}^d$ with characteristic length ℓ, so $k_x(x, x') \le \exp(-\|x - x'\|/\ell)$ up to a kernel-dependent constant. Let $X = \{x_1, \ldots, x_N\} \subset \Omega$ be the design and consider an ℓ-packing $X_\ell \subseteq X$: a maximal subset with pairwise distance $\ge \ell$. Standard volume-comparison gives $|X_\ell| \le (\mathrm{diam}(\Omega)/\ell)^d$.

Effective-rank bound. The off-packing residuals satisfy $|k_x(x_i, x_j)| \ge \exp(-1)$ for x_i, x_j inside the same packing cell, so the kernel matrix $\boldsymbol{K}_x$ is within $O(1)$ of a block-constant matrix with $|X_\ell|$ blocks. Up to a fixed constant, $\boldsymbol{K}_x$ has effective rank $N_{\text{eff}} := |X_\ell| \le \min\{N, (\mathrm{diam}(\Omega)/\ell)^d\}$, and only the top N_{eff} eigenmodes carry signal.

Fisher information. Reusing the eigenmode decomposition of the cr_bound proof, the Fisher information for ρ is $I_N(\rho) = \sum_i I_i(\rho)$ with I_i proportional to the SNR u_i of mode i. Modes outside the packing have $u_i = O(\exp(-2))$ and contribute $O(1)$ in total; the sum is dominated by the N_{eff} in-packing modes:

$$I_N(\rho) \;\le\; N_{\text{eff}}\,\frac{1+\rho^2}{(1-\rho^2)^2} \;\le\; \min\big\{N, (\mathrm{diam}(\Omega)/\ell)^d\big\}\,\frac{1+\rho^2}{(1-\rho^2)^2}.$$

The CR bound then yields $\mathrm{Var}[\hat{\rho}] \ge (1-\rho^2)^2/[N_{\text{eff}}(1+\rho^2)]$: the GP identifies ρ from at most N_{eff} effective paired observations.