On Basis Function Selection for Sparse Gaussian Process Regression

Marnix Van Soom
Vrije Universiteit Brussel
marnix@ai.vub.ac.be

Ivan De Boi
Universiteit Antwerpen
ivan.deboi@uantwerpen.be

Abstract

Sparse Gaussian processes achieve $\mathcal{O}(N)$ inference by replacing the kernel with an appropriate expansion in a fixed basis $\{\phi_j\}$ on the input space. Given a compute budget $M \ll N$, practitioners conventionally *truncate* the basis to its first M entries. Nothing in the formalism, however, prevents one from *selecting* only those M basis functions that matter for the data at hand. This would avoid spending budget on basis functions where there is no signal, but it requires a criterion for ranking the candidates. We propose three such criteria derived from an information-theoretic view of the basis-function selection problem. Each criterion matches a different state of knowledge at selection time: (no data), (no prior), and an (in-between) state. We then study the performance of truncation vs. selection strategies on six UCI regression benchmarks across three basis families: Hilbert-space Gaussian processes (HSGP), variational Fourier features (VFF), and variational inducing spherical harmonics (VISH). We observe that the (no data) criterion is a safe default, matching or improving on truncation for HSGP, VFF, and VISH, with substantial gains for VISH and improvements over a recently developed selection heuristic for that basis family. The data-aware (no-prior) and (in-between) criteria provide substantial gains over truncation specifically for HSGP, which is the most broadly used of the three families in practice.

1 Introduction

Gaussian processes [17] are a powerful class of models for regression. Given N noisy observations $\mathbf{y} = (y_1, \ldots, y_N)$ at inputs $X = (x_1, \ldots, x_N) \in \mathbb{R}^{N \times D}$, one models the response through a latent function f with the prior $f \sim \mathcal{GP}(0, k_\theta)$, and learns the kernel hyperparameters θ from the data. Vanilla inference is, however, $\mathcal{O}(N^3)$ in time because of the kernel-matrix solve, which is prohibitive for the moderately large N encountered in practice.

Sparse GPs [12] avoid this cost by replacing the kernel with an expansion in a basis $\{\phi_j\}$ on the input space, informally written as

$$f(x) \;=\; \sum_{j=1}^{\infty} w_j \, \phi_j(x), \qquad w \sim \mathcal{N}\big(0, \Sigma(\theta)\big). \tag{1}$$

The sparse-GP method fixes both the basis $\{\phi_j\}$ and the covariance $\Sigma(\theta)$ on the coefficients, and typically also requires additional user choices, such as the input domain on which the GP is approximated. One such choice stands out: the number $M \ll N$ of basis functions actually used for inference. Computing the marginal likelihood and the predictive mean scales as $\mathcal{O}(NM^2 + M^p)$, with p depending on the method, so M is essentially the user's compute budget. The conventional way to spend that budget is to *truncate* the expansion in Equation (1) to its first M entries, in some ordering of the index j.

Preprint.

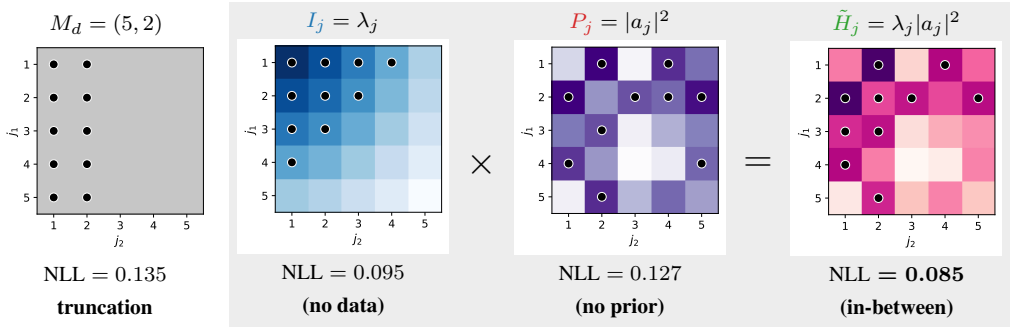

Figure 1: The four basis-function selection strategies we consider in the paper, illustrated on a toy data set. From left to right: truncation $M_d = (5, 2)$, the (no data) criterion $I_j = \lambda_j$, the (no prior) criterion $P_j = |a_j|^2$, and the (in-between) criterion $\tilde{H}_j = \lambda_j |a_j|^2$, computed as the pointwise product of the previous two. Selected basis functions are marked by black dots, and the test NLL is reported underneath each panel. Heatmap colours indicate the strength of the criterion (e.g. I_{j_1,j_2} in panel 2) at each grid point. On this example $\tilde{H}_j$ performs best by combining the low-frequency basis functions favoured by I_j with the higher-energy ones that P_j alone prefers.

For example, take a regression problem in $D = 2$ input dimensions and apply the Hilbert-space sparse-GP method, which is the most common of the three methods considered here (see Section 2 for the construction). The expansion in Equation (1) then takes the form

$$f(x_1, x_2) \;=\; \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} w_{j_1,j_2}\, \phi_{j_1,j_2}(x_1, x_2), \qquad w_{j_1,j_2} \sim \mathcal{N}\big(0, \lambda_{j_1,j_2}(\theta)\big), \tag{2}$$

where the basis function ϕ_{j_1,j_2} is a product of two one-dimensional sinusoids, indexed by per-axis frequencies (j_1, j_2), and the prior variance $\lambda_{j_1,j_2}(\theta)$ is the kernel's spectral density evaluated at the frequency corresponding to (j_1, j_2). Suppose now that our compute budget allows only $M = 10$ basis functions out of the infinite expansion: which ten pairs (j_1, j_2) should we keep, that is, which frequencies should we model?

Figure 1 (left panel) shows one such allocation on a toy data set: the choice $M_d = (5, 2)$ arranges its basis functions as a rectangle in the (j_1, j_2) grid, with the test negative log-likelihood (NLL) reported underneath the panel. Such a rectangular allocation is the convention in available HSGP implementations: the user picks a count M_d per axis,[1] and the method then keeps every basis function with $j_d \le M_d$, giving $M = \prod_{d=1}^{D} M_d$ basis functions in total. This rule, however, is not unique even at fixed M: a budget of $M = 10$ admits the allocations $(5, 2)$, $(2, 5)$, $(10, 1)$, and any other partition with the same product, and the user is left to pick by heuristics and trial and error.

This non-uniqueness in truncation points to a deeper fact: nothing in the sparse-GP formalism actually requires us to truncate sequentially at all. We can instead rank the candidates by any other criterion and keep the top M by that criterion. We call this *selection*, in contrast with the *truncation* baseline. The question is then which criterion to choose, and whether a cheap one can offer gains over truncation.

This paper proposes three such criteria, derived from an information-theoretic view of the problem. They are shown in the grey panels of Figure 1 alongside the truncation baseline. Each criterion produces a scalar value (the score) at every candidate index $j = (j_1, j_2)$, and selection keeps the top M candidates by that value.

We show in Section 3 that the criteria correspond to different states of knowledge at selection time. The (no data) criterion $I_j = \lambda_j(\theta)$ ranks candidates by the kernel-prior weight alone (panel 2). The (no prior) criterion $P_j = |a_j|^2$ ranks by the squared data projection $a_j := \phi_j(X)^\top y$ alone (panel 3). The (in-between) criterion $\tilde{H}_j = \lambda_j |a_j|^2$ uses both factors (panel 4). Black dots mark the selected top-$M = 10$ basis functions, and each criterion picks a different ten. The resulting fit, and therefore the test NLL reported under each panel, varies with the choice of criterion.

[1] For example, PyMC's `pymc.gp.HSGP(m=[`$M_1, \ldots, M_D$`])` and NumPyro's `hsgp_squared_exponential(m=[...])` both take a per-dimension list [24, 19], while the brms R package exposes a single k across all axes via `gp(..., k=...)`.

HSGP truncation can only carve rectangular subsets out of the candidate grid. The three selection criteria are not bound by that shape. Even without looking at the data, the (no data) criterion I_j already departs from the rectangle and improves substantially on truncation. On this toy example, the (in-between) criterion $\tilde{H}_j$ does best, notably better than the rectangular truncation.

We shall see that this pattern does not hold in general. We test whether selection improves over truncation on six UCI regression benchmarks across three sparse-GP methods: HSGP, variational Fourier features (VFF), and variational inducing spherical harmonics (VISH). The three methods are different but related, each with its own basis ϕ_j and prior variance $\lambda_j(\theta)$. Ranking by the (no data) criterion I_j is a safe default that matches or improves on truncation for all three methods, with the largest gains on VISH where it also outperforms the recently proposed phase-truncation heuristic of Eleftheriadis et al. [6]. The data-aware (no prior) criterion P_j and (in-between) criterion $\tilde{H}_j$ offer substantial gains over truncation specifically for HSGP, the most broadly used of the three methods in practice.

The rest of the paper develops these ideas. Sections 2 and 3 fix notation and derive the three criteria. Sections 4–6 report the per-method experiments. Section 7 places the criteria in context, and Section 8 discusses what we learned.

2 Background

We consider the standard Gaussian process regression setting: a Gaussian process $f \sim \mathcal{GP}(0, k_\theta)$ on an input space $\mathcal{X} \ni x = (x_1, \ldots, x_D)$ observed at N inputs $X = (x_1, \ldots, x_N)^\top$ with additive Gaussian noise $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The prior mean is zero throughout this paper, with any mean shift absorbed by standardising the response in pre-processing. Exact GP inference scales as $O(N^3)$ in compute and is infeasible at the dataset sizes we typically care about. We therefore work with sparse GPs throughout. The three methods we consider (HSGP, VFF, and VISH) each specify the sparse GP through a basis-function expansion of f.

Each method supplies an infinite basis-function expansion of the form Equation (1).[2] The index j is in general a multi-index whose structure is set by the method. In practice we keep only M of these basis functions, either by truncation (which keeps the first M in the natural ordering of j) or by selection (which scores a finite candidate grid $\{\phi_j\}_{j=1}^J$ and keeps the top M by score). The *candidate grid size* J is therefore needed only for selection, and in general $M \ll J \ll N$ when N is large. The selection problem itself is to choose a subset $\mathcal{M} \subset \{1, \ldots, J\}$ of size M, and to fit θ on that subset. The three methods specialise the basis and the prior as follows.

Hilbert-space Gaussian processes (HSGP) [24, 19] use Dirichlet Laplacian eigenfunctions on a box $[-L, L]^D$, indexed by $j = (j_1, \ldots, j_D)$, with prior variance $\lambda_j(\theta) = S_\theta(\omega_j)$ at the Laplacian frequency vector $\omega_j = (\pi j_1/2L, \ldots, \pi j_D/2L)$, and are the most broadly used of the three methods in practice. Details in Appendix C.

Variational Fourier features (VFF) [7] use an additive model $f(x) = \sum_{d=1}^D f_d(x_d)$ where each f_d is expanded in windowed cosine and sine features at per-axis frequencies j_d. Candidates are indexed by $j = (d, j_d)$, with prior variance $\lambda_j(\theta) \approx S_\theta(\omega_{j_d})/T$ the kernel's spectral density at the harmonic frequency on the per-axis box of length T. Details in Appendix D.

Variational inducing spherical harmonics (VISH) [5, 6] use spherical harmonics on S^{D-1} for inputs lifted to the unit sphere by $z_n = (x_n, 1)/\|(x_n, 1)\|$, indexed by $j = (\ell, k)$ for degree ℓ and orientation k, with prior variance $\lambda_j(\theta) = \lambda_\ell(\theta)$ shared across orientations within a degree by Funk–Hecke. We use the order-1 arc-cosine kernel of Cho and Saul [3] throughout. Details in Appendix E.

For each method we follow the authors' recommended training procedure: the closed-form marginal likelihood for HSGP, and the closed-form collapsed Titsias bound for VFF and VISH. We reproduce the published truncation baselines as faithfully as possible, with the full experimental protocol in Appendix G.

[2] The basis-function expansion is one of two equivalent ways to specify a sparse GP. The other goes through inducing variables and a variational distribution [12]. The two views agree at the level of the per-basis-function prior weight $\lambda_j(\theta)$ and the data projection a_j that our criteria depend on. Appendix A discusses the connection.

3 The selection criteria

Each candidate index j identifies a basis-function coefficient u_j in the expansion of Equation (1), with marginal prior $u_j \sim \mathcal{N}(0, \lambda_j(\theta))$ in all three methods (Appendix A). We want to know whether u_j is worth keeping in $\mathcal{M}$. The natural quantity is how much observing the data y changes the belief about u_j, measured as the Kullback–Leibler divergence from posterior to prior:

$$H_j(\theta, y) \;:=\; D_{\mathrm{KL}}\big(p(u_j \mid y) \,\|\, p(u_j)\big). \tag{3}$$

This is the per-basis-function information gain: the same D_{KL} objective that Rasmussen and Williams [18] use for placing inducing points and Seeger et al. [21], Krause et al. [9] use for sensor selection, applied here to the discrete index j that enumerates a fixed candidate basis. A coefficient u_j with large H_j is one whose belief is shifted strongly by the data, so a large H_j flags a candidate worth keeping. Since $H_j \geq 0$ we use it directly as a non-negative scoring rule for the basis-function selection problem.

The prior $p(u_j) = \mathcal{N}(0, \lambda_j(\theta))$ is univariate Gaussian by construction, and the marginal posterior $p(u_j \mid y)$ is univariate Gaussian by conjugacy in Bayesian linear regression on the selected basis. Equation (3) therefore has a closed form. With the per-basis-function data projection and signal-to-noise ratio

$$a_j \;:=\; \phi_j(X)^\top y, \qquad \rho_j \;:=\; \frac{\lambda_j(\theta)\,\|\phi_j(X)\|^2}{\sigma^2}, \tag{4}$$

this closed form is

$$H_j(\theta, y) \;=\; \tfrac{1}{2}\Big[\log(1+\rho_j) - \tfrac{\rho_j}{1+\rho_j}\Big] \;+\; \frac{\lambda_j(\theta)\,|a_j|^2}{2\,\sigma^4\,(1+\rho_j)^2}, \tag{5}$$

under the empirical-decoupling assumption $\Phi^\top\Phi \approx \mathrm{diag}(\|\phi_j(X)\|^2)$, the standard regime in the basis-function case (Appendix B). The first bracket is a variance-shrinkage term that is independent of the data. The second bracket is the data-driven term. Evaluating Equation (5) requires both the noise scale σ^2 and the kernel eigenvalues $\lambda_j(\theta)$, neither of which is fit yet at selection time.

Three limits, three criteria Three limits of H_j correspond to three states of knowledge at selection time about how the prior weight and the data interact, and each gives a closed-form ranking criterion on the candidate basis.

(i) The no-data limit: the eigenvalue criterion I_j. Under the prior, a_j is mean-zero, and a short calculation (Appendix B) gives the data-averaged information gain

$$\mathbb{E}_y\big[H_j(\theta, y)\big] \;=\; \tfrac{1}{2}\log(1+\rho_j), \tag{6}$$

monotone in ρ_j and therefore in the prior weight $\lambda_j(\theta)$ at fixed θ and σ^2. Ranking by Equation (6) therefore reduces to ranking by

$$I_j(\theta) \;:=\; \lambda_j(\theta). \tag{7}$$

At fixed θ, I_j ranks candidates by prior weight alone. It is the closest per-basis-function analogue of conventional truncation, since for stationary kernels the spectral density decays radially and I_j keeps the lowest-frequency basis functions first. The two need not agree, however: practitioner-default truncation enforces a structural shape (rectangular cube, harmonic shells, contiguous frequencies), while I_j ranks every candidate freely (Figure 1, panels 1 and 2).

(ii) The no-prior limit: the data-energy criterion P_j. Dropping the kernel factor in Equation (5) leaves the data-driven piece, $|\phi_j(X)^\top y|^2/[2\sigma^4(1+\rho_j)^2]$. At fixed θ and σ^2 the prefactors are basis-function-independent up to $(1+\rho_j)^2$, which itself is a function of λ_j. Ignoring that residual prior dependence gives the data-energy criterion

$$P_j(y) \;:=\; |\phi_j(X)^\top y|^2, \tag{8}$$

the squared projection of the data onto ϕ_j. A second, more principled, route arrives at the same ranking: if σ^2 and λ_j are jointly estimated from the data and substituted into the full Equation (5), the resulting per-basis-function score is rank-equivalent to P_j (Appendix B). The fully data-driven H_j therefore collapses to the data-only ordering.

Algorithm 1 Selecting basis functions by score

1: Enumerate $J \gg M$ candidate basis functions $\{\phi_j\}_{j=1}^{J}$ from the chosen basis family.
2: Choose a starting point θ_0 for the kernel hyperparameters.
3: Compute the kernel weights $\lambda_j(\theta_0)$ for all candidates in $\mathcal{O}(J)$ time.
4: Compute the data projections $a_j := \phi_j(X)^\top y$ for all candidates in $\mathcal{O}(NJ)$ time.
5: Choose one ranking score:

$$I_j(\theta_0) = \lambda_j(\theta_0), \qquad P_j(y) = |a_j|^2, \qquad \tilde{H}_j(\theta_0, y) = \lambda_j(\theta_0)|a_j|^2.$$

6: Rank candidates greedily in descending order by the chosen score in $\mathcal{O}(J)$ time.
7: Break ties by axis order, with axis 1 most important.
8: Return the top-M indices.

(iii) The weak-observation limit: the criterion $\tilde{H}_j$. Selection happens before the kernel hyperparameters are fit, so the practical setting is one of weak observation: the score is evaluated at a starting point θ_0 rather than at a maximum-likelihood estimate. Taking $\rho_j \to 0$ in Equation (5),

$$H_j(\theta, y) \;=\; \frac{\lambda_j(\theta)\,|a_j|^2}{2\,\sigma^4} \;+\; O(\rho_j^2), \tag{9}$$

the variance-shrinkage bracket drops to second order, only the data-driven term survives, and the noise scale σ^2 factors out of the ranking. Define

$$\tilde{H}_j(\theta, y) \;:=\; \lambda_j(\theta) \;\cdot\; \left|\phi_j(X)^\top y\right|^2. \tag{10}$$

The weak-observation criterion retains both the kernel and the data factor, while σ^2 drops out: ranking depends only on the basis ϕ_j, the kernel eigenvalues at θ_0, and the data y.

A single pass over the data The data factor $|a_j|^2$ in P_j and $\tilde{H}_j$ has a periodogram structure: it is a windowed periodogram of y against $\{\phi_j\}_{j=1}^{J}$, and the full vector $\Phi^\top y$ is a identical to a $\mathcal{O}(NJ)$ transform of the data over the candidate basis. In practice the user should pick J as large as the candidate-precomputation budget allows: a larger J widens the frequency or degree coverage of the candidate basis before any are dropped. Since the basis ϕ_j never depends on θ in any of the three methods, this transform is computed once, before any hyperparameter fit, and yields P_j and $\tilde{H}_j$ at any starting point θ_0. Its cost is small relative to a single hyperparameter-fit step (which itself involves an $\mathcal{O}(NM^2)$ linear solve), so it is effectively amortised by the downstream fit.

Algorithm Algorithm 1 summarises the selection step for any of the three criteria as a single rank-and-truncate pass. It is important to note that the proposed criteria are heuristics. None of them is the canonical H_j at the eventual fitted hyperparameters, and the greedy rank-and-truncate approach comes with no theoretical guarantee on downstream test performance. Whether they nonetheless do useful work is an empirical question, which the next three sections take up.

4 Experiment I: Hilbert-space Gaussian processes

We sweep six UCI regression benchmarks (concrete, energy, kin8nm, power, yacht, airfoil) over $M \in \{16, 32, 64, 128, 256\}$ across 10 random 90:10 train/test splits with a Matérn-5/2 ARD kernel. Construction, candidate-basis size, fit objective, and the non-uniform truncation baseline are in Appendix C (and the shared dataset and optimisation protocol in Appendix G).

Findings Both data-aware criteria improve on the (no data) I_j on most cells, with P_j slightly stronger on average and $\tilde{H}_j$ sitting between I_j and P_j in many cells. The improvement is largest at small M, where allocating budget to the basis functions the data actually activates pays off most, and shrinks toward larger M as the budget catches up with the signal. The non-uniform truncation baseline tracks I_j closely on every cell and is slightly worse on average, an empirical sanity check that the per-basis-function eigenvalue ordering is at least as good as the structural product rule. The HSGP candidate grid is anisotropic by construction with active axes unknown before fitting, and the

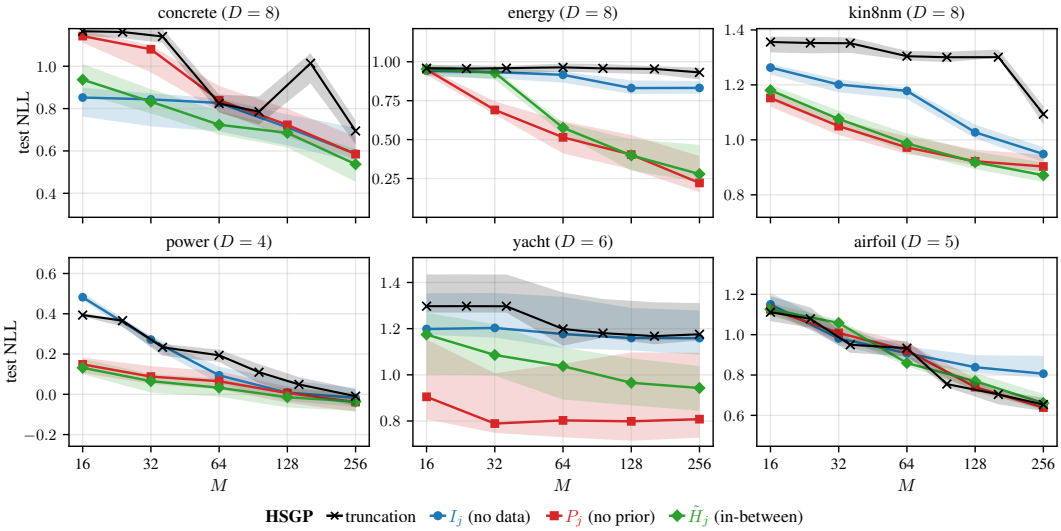

Figure 2: Hilbert-space Gaussian processes (HSGP) on six UCI regression benchmarks. Median test NLL versus compute budget M across 10 random 90:10 train/test splits, with interquartile bands. The data-aware (in-between) $\tilde{H}_j$ and (no prior) P_j improve on the (no data) I_j on most cells, with P_j slightly stronger on average. The non-uniform per-axis truncation baseline tracks I_j within noise, slightly worse on average. Truncation x-axis at the nearest reachable budget under $M = \prod_d M_d$ (Appendix C).

data projection $|a_j|^2$ gives a cheap early signal that the data-aware criteria pick up directly while I_j has to wait for the hyperparameter fit to discover it.

5 Experiment II: Variational Fourier features

We use the same six UCI benchmarks and split protocol as for HSGP, with a per-axis Matérn-$5/2$ kernel, the candidate basis flattened across axes and ranked globally, and the closed-form collapsed Titsias bound fit by L-BFGS-B from the unit starting point. Construction, the periodic-extension diagonalisation, and the per-axis truncation baseline are in Appendix D.

Findings Figure 3 shows that $\tilde{H}_j$ is roughly flat against I_j across most cells, while P_j is reliably worse than both. The per-axis truncation baseline sits on top of I_j at essentially every cell: the two rules pick almost identical basis function sets, because the per-axis eigenvalue order is already the lowest-frequency-first order that truncation enforces. On smooth UCI regressions the data projection $|a_j|^2$ mostly adds noise to an ordering that the eigenvalues already get right.

6 Experiment III: Variational inducing spherical harmonics

We use the same six UCI benchmarks and split protocol as for HSGP, with the order-1 arc-cosine kernel [3] on the lifted inputs and the gpfy reference implementation [6] for the spherical-harmonic basis and its Funk–Hecke spectrum. Only the basis-selection rule changes between curves, and we compare against two practitioner-default truncation rules from the literature: the cumulative-shell rule of Dutordoir et al. [5] and the phase-truncation rule of Eleftheriadis et al. [6]. Construction, the closed-form arc-cosine zero pattern, and the two truncation baselines are in Appendix E.

Findings Against I_j, the data-aware criteria do not help on this family. The (no data) I_j already captures the right structural prior: spherical-harmonic eigenvalues drop fast with degree, and on smooth UCI regressions the signal energy is concentrated at low degree, so reordering by data energy mostly perturbs an already-good ranking. The cumulative-shell truncation curve, however, behaves more dramatically: at cumulative-shell budgets that end on an even degree it selects bit-for-bit the

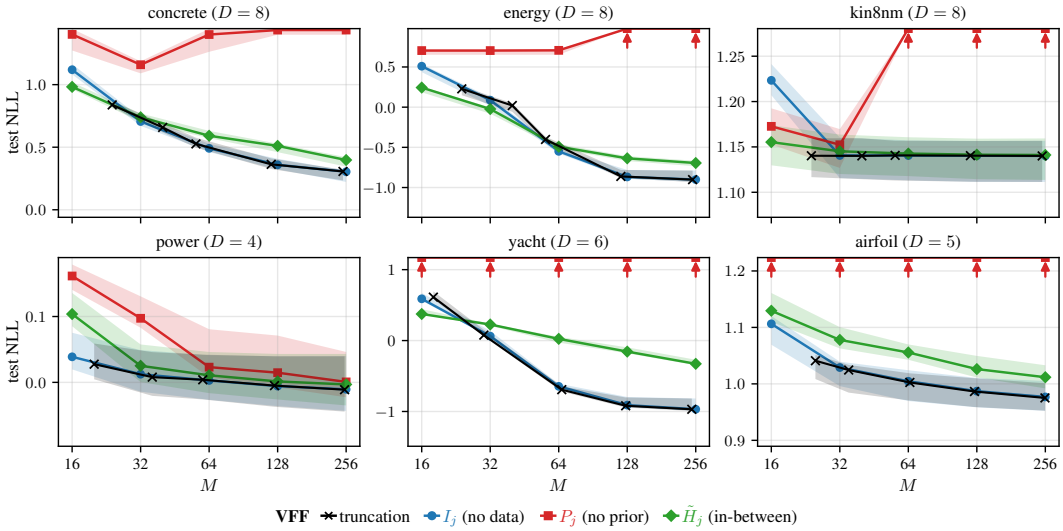

Figure 3: Variational Fourier features (VFF) on the same setup as Figure 2 (Matérn-$5/2$ per axis, closed-form collapsed Titsias bound). The data-aware (in-between) $\tilde{H}_j$ and (no prior) P_j neither help nor hurt against the (no data) I_j at larger budgets, with $\tilde{H}_j$ slightly behind on average. The conventional VFF per-axis truncation baseline sits on top of I_j at every cell. Truncation x-axis at the nearest reachable budget under $M = D(2M_\star - 1)$ (Appendix D).

same basis functions as I_j, but whenever the cutoff $L^\star$ lands on an odd degree $\ell \geq 3$ the curve jumps upward. The cause is a closed-form property of the order-1 arc-cosine kernel: its Funk–Hecke expansion has $\mu_\ell = 0$ for all odd $\ell \geq 3$ [1, Appendix D.2], independent of the sphere dimension D, so filling an odd shell spends budget on coefficients the kernel forces to zero variance, and the (no data) ranking skips those shells automatically. Against the phase-truncation baseline, every score-based criterion is consistently better, with the gap reaching one to two orders of magnitude in test NLL at small M on energy and power: the single global $m^\star$ knob cannot adapt when the data favour more basis functions at one degree than another, and it inherits the same odd-shell trap. The phase-truncation sweep also takes about 10.5 seconds per fit versus 1.4 seconds for the score-selected VISH sweep, with about $18\times$ as many L-BFGS-B iterations, because the phase variables are optimised inside the fit whereas I_j fixes a subset before fitting.

7 Related work

Active-set and information-theoretic selection in Gaussian processes Sparse-GP active-set methods have long used greedy and information-theoretic criteria to decide which scalar linear functionals of f enter the approximation. The informative vector machine of Lawrence et al. [10], fast forward selection for sparse GP regression [21], sparse online GPs [4], and sparse greedy GP regression [22] all construct a small active set this way, and the same KL information gain $D_{\mathrm{KL}}(p(u_j \mid y) \,\|\, p(u_j))$ that we use for H_j underlies several of these scores [18, 9]. More recent work continues this line by placing Bayesian or task-specific structure on inducing locations and their number, for example through point-process priors [26], Bayesian inference over inducing locations [20], or quality-diversity allocation for Bayesian optimisation [14]. For continuous inducing-point locations $z \in \mathcal{X}$ the criterion is differentiable and is typically optimised jointly with the variational posterior [23], but all of these methods select, infer, or allocate data-indexed inducing inputs as a sub-routine of inference itself. Our setting is different. The basis families we consider already define a fixed candidate set of linear functionals (HSGP eigenfunctions, VFF Fourier features, VISH spherical harmonics), and the question is only how to spend a given budget M inside that set. Once this candidate basis is fixed, each criterion is a scalar function of the prior weight λ_j and the data projection $a_j = \phi_j(X)^\top y$, and all candidates can be ranked in one pass before any posterior fit. The contribution is therefore not a new active-set sparse GP algorithm, but a basis-index selection rule that can replace the default truncation rule inside existing basis-function sparse GP constructions.

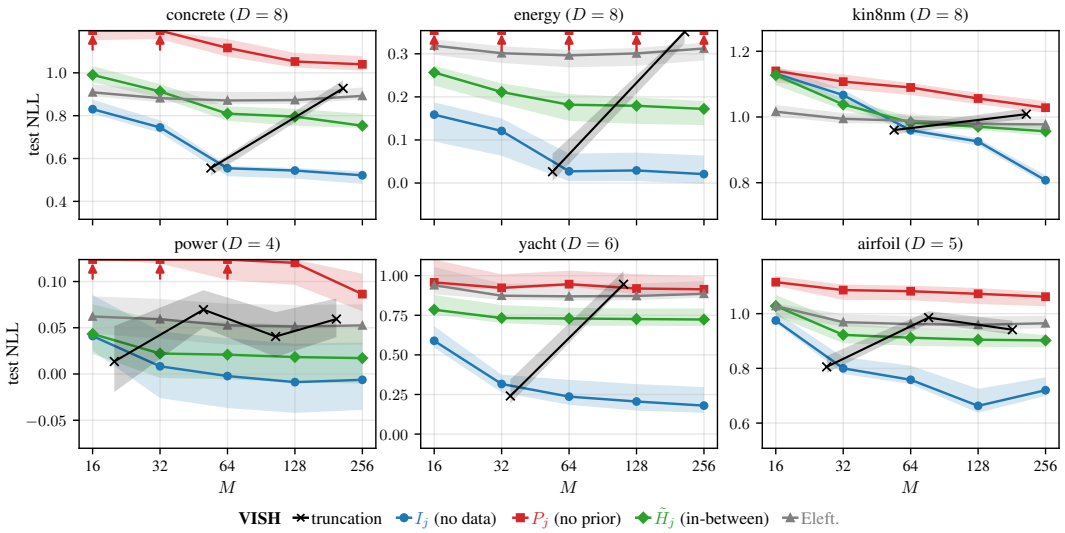

Figure 4: Variational inducing spherical harmonics (VISH) on the same setup as Figure 2 (arc-cosine order-1 kernel, spherical harmonics on S^D). Against the (no data) I_j, the data-aware (in-between) $\tilde{H}_j$ does not improve on most cells. The cumulative-shell truncation curve jumps upward whenever the cutoff $L^\star$ lands on an odd degree (see body). Against the Eleftheriadis et al. [6] phase-truncation baseline, every score-based criterion is consistently better, often by orders of magnitude at small M. Both truncation x-axes at the nearest reachable budget under each rule's structural constraint (Appendix E).

Matching pursuit and correlation screening The (no prior) score $P_j = |a_j|^2$ is the same correlation-screening step that initialises matching pursuit and its orthogonal variant [13, 15], in which dictionary atoms are picked by their alignment with the signal or current residual. Two differences matter. Matching pursuit is iterative: after an atom is selected the residual is updated, and in orthogonal matching pursuit the selected atoms are re-orthogonalised. Our selection step is a non-iterative rank-and-truncate pass that fixes the GP basis once, before any fit. The GP setting also supplies a prior variance λ_j for each candidate coefficient, so the (no data) $I_j = \lambda_j$ and (in-between) $\tilde{H}_j = \lambda_j |a_j|^2$ are not pure correlation-screening rules. They combine the observed alignment with the kernel-implied plausibility of the candidate, which is what makes the same selection idea meaningful across HSGP, VFF, and VISH, whose basis indices have different geometries but all carry GP prior weights.

Spectral and eigenfunction feature methods A different line of work changes the feature construction itself. Sparse-spectrum GPs choose a small set of spectral frequencies and learn their locations as model parameters [11]. Variational orthogonal features construct stationary-kernel features for cheaper variational inference [2], and data-dependent eigenfunction methods select eigenfunctions of an empirical Gram matrix by evidence maximisation [16]. These ask which feature family or spectral approximation should define the approximation. Our setting is narrower but complementary: the feature family is fixed by a published sparse-GP construction, and we only choose the subset of its candidate indices.

8 Discussion

The work done here is essentially a preliminary investigation of a gap in the sparse-GP literature. The three methods we considered all recommend truncation as the default way to spend the basis budget. Yet for a fixed compute budget M (the realistic inference setting), there is freedom in choosing *which* basis functions to actually use, and the truncation default does not exercise it. Asking whether the candidate basis should be selected instead turns out to be cheap enough to test, and informative enough to separate from a broad benchmark of sparse-GP approximations. The investigation has taught us two things.

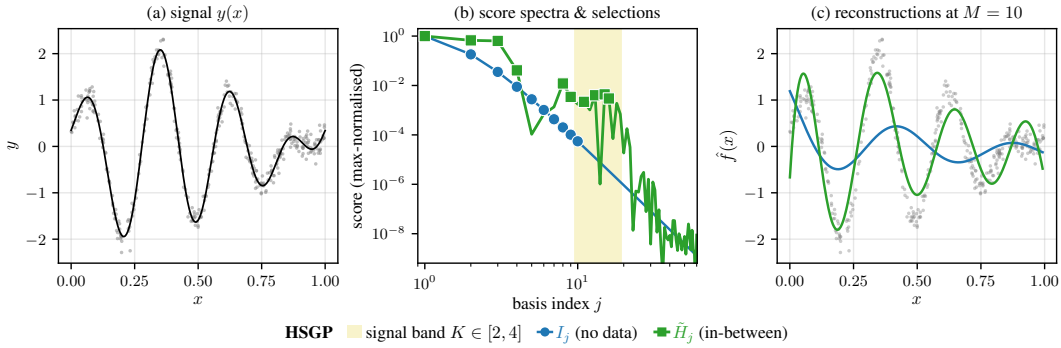

Figure 5: Toy bandpass example (details in Appendix F). (a) Signal in the integer-frequency band $K \in [2, 4]$ plus noise. (b) Score spectra: the (no data) I_j is monotone low-pass and picks the lowest frequencies (circles), while the (in-between) $\tilde{H}_j$ rises inside the band and picks there (squares). (c) HSGP fits at $M = 10$: I_j over-smooths through the band, while $\tilde{H}_j$ tracks the wiggles.

The first is that the (no data) criterion I_j is a strong default: ranking every candidate freely by its kernel-prior weight is at least as good as the standard truncation rule for each method. On VISH it is indeed much better than either of the two practitioner-default truncation rules we compared against: an odd-shell zero pattern in the order-1 arc-cosine kernel kills entire structural blocks that the standard truncation rules waste budget on, while I_j avoids them automatically (Section 6). The same simple ranking also defeats the phase-truncation baseline of Eleftheriadis et al. [6], with substantial speedups in addition to the accuracy gain (Section 6).

The second is that the data-aware (no prior) and (in-between) criteria are method-specific. On HSGP the candidate grid (a high-dimensional version of the rectangle in Figure 1) has unknown active axes before fitting, so the data projection $|a_j|^2$ is the cheapest probe of where the signal lives, and both P_j and $\tilde{H}_j$ give an essentially free improvement over truncation. On VFF and VISH the basis is naturally ordered to favour low-frequency or low-degree functions, which is what smooth UCI regressions need, and I_j alone exploits this. The data-aware criteria are most likely to help when the data has structure that low-frequency truncation misses, as in the toy bandpass example of Figure 5.

Limitations and future work The eigenvalue-dependent criteria I_j and $\tilde{H}_j$ evaluate λ_j at a single guessed starting point θ_0 rather than integrating over plausible hyperparameters, and the closed-form H_j is exact only under the empirical-decoupling assumption $\Phi^\top \Phi \approx \mathrm{diag}(\|\phi_j(X)\|^2)$. The three methods also differ in how their basis is constructed (HSGP native multi-dimensional, VFF additive across axes, VISH native multi-dimensional on the lifted sphere), so cross-method differences in selection behaviour can reflect this construction confound as well as the criteria. The empirics are preliminary, with six UCI regressions at five budgets per method and test RMSE corroborating the test-NLL pattern (Appendix H). Whether the patterns generalise to other kernels, basis families, or larger-scale regressions remains for future work.

Conclusion This paper addressed a gap in the basis-function sparse-GP literature: the practitioner has a fixed compute budget M to spend across the candidate basis, but the default answer is structural truncation. We asked whether a simple one-pass selection criterion can do better, and the answer turns out to be yes, with the size of the win depending on the method. The (no data) I_j ties or beats truncation everywhere and substantially improves over it on VISH, while the data-aware (no prior) P_j and (in-between) $\tilde{H}_j$ criteria further improve on HSGP, which is the most adopted method in practice.

References

[1] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. Appendix D.2 gives the closed-form Funk–Hecke multipliers for the order-1 arc-cosine (ReLU) kernel, with the parity property $\mu_\ell = 0$ for odd $\ell \geq 3$.

[2] David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of Sparse Variational Inference in Gaussian Processes Regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020. ISSN 1533-7928.

[3] Youngmin Cho and Lawrence Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[4] Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14 (3):641–668, 2002.

[5] Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse Gaussian Processes with Spherical Harmonic Features. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2793–2802. PMLR, November 2020.

[6] Stefanos Eleftheriadis, Dominic Richards, and James Hensman. Sparse Gaussian Processes with Spherical Harmonic Features Revisited, March 2023.

[7] James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for gaussian processes. *The Journal of Machine Learning Research*, 18(1):5537–5588, 2017.

[8] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. `https://archive.ics.uci.edu`, 2023.

[9] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. In *Journal of Machine Learning Research*, volume 9, pages 235–284, 2008.

[10] Neil D Lawrence, Matthias Seeger, and Ralf Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems*, volume 15, 2002.

[11] Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R Figueiras-Vidal. Sparse Spectrum Gaussian Process Regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

[12] Felix Leibfried, Vincent Dutordoir, S. T. John, and Nicolas Durrande. A Tutorial on Sparse Gaussian Processes and Variational Inference, December 2022.

[13] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[14] Henry B. Moss, Sebastian W. Ober, and Victor Picheny. Inducing point allocation for sparse Gaussian processes in high-throughput Bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5213–5230. PMLR, 2023.

[15] Yagyensh C Pati, Ramin Rezaiifar, and Perinkulam S Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

[16] Yuan Qi, Ahmed H Abdel-Gawad, and Thomas P Minka. Sparse-posterior Gaussian processes for general likelihoods. In *Uncertainty in Artificial Intelligence*, 2010.

[17] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9.

[18] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Section 8.3 introduces the per-coefficient KL information gain as a criterion for active inducing-set selection in sparse GPs.

[19] Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *arXiv:2004.11408 [stat]*, April 2020.

[20] Simone Rossi, Markus Heinonen, Edwin Bonilla, Zheyang Shen, and Maurizio Filippone. Sparse Gaussian processes revisited: Bayesian approaches to inducing-variable approximations. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1837–1845. PMLR, 2021.

[21] Matthias W. Seeger, Christopher K. I. Williams, and Neil D. Lawrence. Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. In *International Workshop on Artificial Intelligence and Statistics*, pages 254–261. PMLR, January 2003.

[22] Alex J. Smola and Peter L. Bartlett. Sparse greedy Gaussian process regression. In *Advances in Neural Information Processing Systems*, volume 13, pages 619–625. MIT Press, 2001.

[23] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, volume 18, 2006.

[24] Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30(2):419–446, March 2020. ISSN 1573-1375. doi: 10.1007/s11222-019-09886-w.

[25] Michalis Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 567–574. PMLR, April 2009.

[26] Anders Kirk Uhrenholt, Valentin Charvet, and Bjørn Sand Jensen. Probabilistic selection of inducing points in sparse Gaussian processes. In *Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1035–1044. PMLR, 2021.

A Sparse GPs and basis function expansions

The main text uses the language of finite basis expansions because that is where the selection problem is most transparent: once a candidate dictionary has been fixed, one has to decide which coefficients to keep. Sparse GPs in the sense of Leibfried et al. [12], however, are slightly broader objects than finite expansions. We spell out the connection here so that the basis-expansion view used throughout the paper is not mistaken for the most general sparse-GP construction.

A GP is already a basis function expansion A Gaussian process $f \sim \mathcal{GP}(0, k)$ on a domain $\mathcal{X}$ admits a Mercer expansion of its kernel,

$$k(x, x') \;=\; \sum_{j=1}^{\infty} \lambda_j \, \phi_j(x) \, \phi_j(x'),$$

in an orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ of $L^2(\mathcal{X}, \mu)$ for a natural measure μ, with non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$. By Karhunen–Loève, a draw $f \sim \mathcal{GP}(0, k)$ is then

$$f(x) \;=\; \sum_{j=1}^{\infty} w_j \, \phi_j(x), \qquad w_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \lambda_j).$$

Thus a GP is already a basis function expansion. What is infinite-dimensional is the number of random coefficients, not the expansion form itself.

The Leibfried sparse GP, in general The general definition of Leibfried et al. [12] does not require Mercer. A sparse GP starts from a GP $f \sim \mathcal{GP}(0, k)$ conditioned on a finite set of inducing variables $\mathbf{u} = (u_1, \ldots, u_M)$, each a scalar linear functional of f, together with some chosen distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}_u, S_{uu})$. Marginalising out gives the sparse-GP marginal

$$f(\cdot) \;\sim\; \mathcal{GP}\Big(\mu(\cdot) + k_{\cdot u}\, K_{uu}^{-1}(\mathbf{m}_u - \mu_u),\; k(\cdot, \cdot') - k_{\cdot u}\, K_{uu}^{-1}\,(K_{uu} - S_{uu})\, K_{uu}^{-1}\, k_{u \cdot'}\Big).$$

The covariance is the prior kernel minus a rank-M correction, not a rank-M kernel itself, and so the random function f retains infinite-dimensional uncertainty around the conditional mean. A sparse GP per Leibfried is therefore not a finite basis function expansion in general.

Mercer connects the two Take the inducing variables to be Mercer coefficients, $u_j := \langle f, \phi_j \rangle$, for the first M Mercer eigenfunctions. By orthonormality and Mercer, $K_{uu} = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$ and $k_{xu}[j] = \lambda_j \phi_j(x)$. Substituting into Leibfried's covariance with $S_{uu} = K_{uu}$, the sparse-GP covariance becomes the tail of the Mercer expansion,

$$\mathrm{Cov}_{\text{sparse}}(f(x), f(x')) \;=\; \sum_{j>M} \lambda_j \, \phi_j(x) \, \phi_j(x'),$$

and the GP itself decomposes as

$$f(x) \;=\; \underbrace{\sum_{j=1}^{M} u_j \, \phi_j(x)}_{\text{rank-}M\text{ basis expansion}} \;+\; \underbrace{h(x)}_{\text{Mercer tail}} \;,$$

with $u_j \sim \mathcal{N}(0, \lambda_j)$ and h a zero-mean GP whose covariance is the tail of the Mercer expansion. The finite basis function expansion view drops the residual h. The full sparse-GP view keeps it. Both are valid sparse-GP constructions, and the difference is the orthogonal complement of $\mathrm{span}\{\phi_j\}_{j=1}^{M}$.

What each family in this paper does HSGP [24, 19] drops the residual: the model is an explicit basis function expansion, with Dirichlet Laplacian eigenfunctions ϕ_j and prior variances $\lambda_j(\theta)$ from the kernel's spectral density, and inference is closed-form Bayesian linear regression on the truncated basis. VFF [7] keeps the residual: the inducing variables are RKHS Fourier projections, and Hensman et al. comment explicitly that the variational form does not discard the orthogonal complement. VISH [5, 6] keeps the residual on the same logic with spherical-harmonic projections. Appendices C–E give the construction of each family.

Why the residual does not affect our criteria The criteria $I_j(\theta)$, $P_j(y)$, $\tilde{H}_j(\theta, y)$ are derived from the per-basis-function KL information gain $H_j(\theta, y) = D_{\mathrm{KL}}\big(p(u_j \mid y) \,\|\, p(u_j)\big)$. The prior $p(u_j) = \mathcal{N}(0, \lambda_j(\theta))$ depends only on the kernel eigenvalue at ϕ_j, not on whether the residual is kept or dropped. The marginal posterior $p(u_j \mid y)$ depends on the data through the projection $a_j = \phi_j(X)^\top y$ and on the noise scale σ^2, and under the empirical-decoupling assumption $\Phi^\top \Phi \approx \mathrm{diag}(c_j)$ takes the same value in both pictures: the residual h contributes only to the unmodelled uncertainty in $f(x)$ at unseen test inputs, not to the marginal posterior on the j-th Mercer coefficient. The criteria, and the selection problem they pose, are therefore identical in both pictures.

Conventions used in the main text Given this equivalence at the criterion level, the main text uses the basis-expansion picture as its default exposition. It poses the selection-versus-truncation question cleanly (which Mercer coefficients to keep), unifies the three families at the model-and-prior level (each picks $\{\phi_j\}$ and $\lambda_j(\theta)$), and does not commit to any inference machinery. The full sparse-GP construction is invoked per family only when reporting the fitting procedure each family uses downstream of selection: the closed-form marginal likelihood for HSGP, and the collapsed Titsias bound for VFF and VISH.

B Derivation of the criteria

We derive the closed form Equation (5), the data-averaged form $\mathbb{E}_y[H_j] = \frac{1}{2}\log(1+\rho_j)$, and the rank-equivalence of the fully data-fit H_j with the no-prior criterion P_j. The same argument extends to arbitrary scalar linear functionals of f (inducing points, Mercer projections, windowed Fourier projections). The basis-function specialisation used in the main text follows.

Closed form for H_j Under the empirical-decoupling assumption $\Phi^\top \Phi \approx \mathrm{diag}(c_j)$ with $c_j := \|\phi_j(X)\|^2$, the marginal posterior on the basis-function coefficient u_j is Gaussian with variance $v_j^{\text{post}} = v_j/(1+\rho_j)$ and mean $\mu_j(y) = \lambda_j a_j/(\sigma^2 + \lambda_j c_j)$, where $v_j = \lambda_j(\theta)$ is the prior variance and $\rho_j = \lambda_j c_j/\sigma^2$ is the per-basis-function signal-to-noise ratio. Plugging into the univariate Gaussian KL gives Equation (5).

Data-averaged form: derivation of I_j Under the prior, $a_j = \phi_j(X)^\top y$ has variance $\mathrm{Var}(a_j) = \sigma^2 c_j(1+\rho_j)$, so $\mathbb{E}_y[|a_j|^2] = \sigma^2 c_j(1+\rho_j)$. Substituting into the data-driven term in Equation (5),

$$\mathbb{E}_y\left[\frac{\lambda_j |a_j|^2}{2\sigma^4(1+\rho_j)^2}\right] = \frac{\lambda_j c_j}{2\sigma^2(1+\rho_j)} = \frac{\rho_j}{2(1+\rho_j)},$$

and adding the variance-shrinkage bracket $\frac{1}{2}[\log(1+\rho_j) - \rho_j/(1+\rho_j)]$ collapses the data-driven and shrinkage pieces:

$$\mathbb{E}_y[H_j(\theta, y)] \;=\; \tfrac{1}{2}\log(1+\rho_j),$$

which is monotone in ρ_j at fixed σ^2, c_j and therefore equivalent in rank to $\lambda_j(\theta)$. This is the criterion I_j of Equation (7).

Rank-equivalence of the fully data-fit H_j with P_j If σ^2 and λ_j are jointly fit by maximum likelihood at fixed j, the optimum gives $\hat{\rho}_j = \max(0, |a_j|^2/(\sigma^2 c_j) - 1)$. Substituting back into Equation (5) reduces H_j to a monotone function of $|a_j|^2$ alone, which is rank-equivalent to $P_j = |a_j|^2$.

C Hilbert-space Gaussian processes (HSGP)

For HSGP [24, 19], the candidate dictionary is fixed by the geometry of a box and a Dirichlet boundary condition, while the kernel enters through spectral weights on that dictionary. The details below give the basis $\{\phi_j\}$, the prior variances $\lambda_j(\theta)$, the downstream fit after selection, and the per-axis truncation baseline reported in Section 4.

The basis on a box Take the input domain to be $\Omega = [-L, L]^D$. The Dirichlet eigenproblem for the Laplacian on Ω,

$$-\nabla^2 \phi_j(x) = \mu_j\, \phi_j(x), \qquad x \in \Omega, \qquad \phi_j\big|_{\partial\Omega} = 0,$$

has eigenfunctions

$$\phi_j(x) = \prod_{d=1}^{D} L^{-1/2} \sin\!\left(\frac{\pi j_d (x_d + L)}{2L}\right), \qquad j = (j_1, \ldots, j_D) \in \mathbb{N}_+^D,$$

with eigenvalues

$$\mu_j = \sum_{d=1}^{D} \left(\frac{\pi j_d}{2L}\right)^2.$$

The $\{\phi_j\}$ are orthonormal in $L^2(\Omega)$ and depend only on Ω and the boundary condition. In particular they do not depend on the kernel hyperparameters θ.

Approximate Mercer interpretation For a stationary kernel k_θ on $\mathbb{R}^D$ with spectral density S_θ, the HSGP construction approximates the kernel by

$$k_\theta(x, x') \;\approx\; \sum_j \lambda_j(\theta)\, \phi_j(x)\, \phi_j(x'), \qquad \lambda_j(\theta) = S_\theta(\omega_j),$$

where $\omega_j = (\pi j_1/2L_1, \ldots, \pi j_D/2L_D)$ is the per-axis Laplacian frequency vector (the box $\Omega = \prod_d [-L_d, L_d]$ is allowed to be anisotropic, and the kernel can be ARD). The approximation comes from restricting to a compact Ω, imposing Dirichlet boundary conditions, and truncating to finitely many modes. It is asymptotically exact as the per-axis box lengths $L_d \to \infty$ for inputs away from $\partial\Omega$ and with sufficient mode count. Throughout this paper we use the box $L_d = 1.2 \cdot \max_i |x_{id}|$ per axis, following the convention of Riutort-Mayol et al. [19].

Inference under Gaussian likelihood Once a subset $\mathcal{M}$ of size M is selected, the model is the finite linear-Gaussian system $y = \Phi u + \varepsilon$ with $\Phi_{nj} = \phi_j(x_n)$, basis-function coefficients $u = (u_j)_{j\in\mathcal{M}} \sim \mathcal{N}(0, \Lambda_\theta)$, $\Lambda_\theta = \mathrm{diag}(\lambda_j(\theta))_{j\in\mathcal{M}}$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$. The marginal likelihood,

$$\log p(y \mid \theta, \mathcal{M}) = -\tfrac{1}{2}\left[\log|\Sigma_y| + y^\top \Sigma_y^{-1} y + N\log 2\pi\right], \qquad \Sigma_y = \Phi\Lambda_\theta\Phi^\top + \sigma^2 I_N,$$

is closed-form, so the downstream fit is L-BFGS-B maximisation of this objective rather than variational inference.

Truncation baseline The conventional HSGP library default fixes per-axis counts $(M_1, \ldots, M_D)$ and keeps every basis function with $j_d \leq M_d$ on every axis, giving a total of $M = \prod_d M_d$ basis functions. For our budget range $M \in [16, 256]$ this rule is coarse-grained: a single uniform $M_d = M_\star$ on every axis only lands on $M = M_\star^D$, which covers at most one or two budgets per dataset. We therefore use a non-uniform extension to obtain a fair comparison curve on the same M-grid as the selection criteria. For each target budget we pick the most-uniform integer factorisation $\prod_d M_d$, breaking ties by smallest standard deviation across axes, and assign the descending-sorted tuple to axes ordered by descending training-input variance. This defines a reproducible non-uniform truncation baseline on the same M-grid as the selection criteria, used in Figure 2.

D Variational Fourier features (VFF)

For VFF [7], the candidate basis is Fourier rather than Laplacian, and the multi-dimensional candidate basis used here is assembled from separate one-dimensional Fourier blocks rather than from their tensor product. We describe the one-dimensional features, the diagonalisation approximation used for scoring, the Gaussian-likelihood objective, and the per-axis truncation baseline used in Section 5.

Fourier features on an interval Take the input domain to be $[a, b]$ with length $T = b - a$, and use the normalised measure $d\mu(x) = dx/T$. Define harmonic frequencies $\omega_m = 2\pi m/T$ for $m = 1, 2, \ldots$ and the orthonormal Fourier features

$$\phi_0(x) = 1, \qquad \phi_{c,m}(x) = \sqrt{2}\cos\big(\omega_m(x-a)\big), \qquad \phi_{s,m}(x) = \sqrt{2}\sin\big(\omega_m(x-a)\big).$$

These are orthonormal in $L^2([a, b], \mu)$ and depend only on the interval and the frequency grid. For the multi-dimensional case we use $f(x) = \sum_{d=1}^{D} f_d(x_d)$ with a separate 1D VFF basis per axis. The total candidate count is $\sum_d M_d$, not the tensor-product $\prod_d M_d$, so the construction scales linearly in D at fixed per-axis count.

Periodic-extension diagonalisation The Fourier features are not exact Mercer eigenfunctions of a stationary kernel on a finite interval: boundary effects prevent perfect diagonalisation. Hensman et al. [7] show that for Matérn kernels the exact K_{uu} is diagonal-plus-low-rank, with the diagonal proportional to $1/S_\theta(\omega_m)$ and rank-one corrections capturing the boundary effects. We adopt the standard periodic-extension approximation, in which the kernel is extended periodically with period T and diagonalises in the Fourier basis with spectral weights

$$\lambda_m(\theta) \;\approx\; \frac{S_\theta(\omega_m)}{T}.$$

The approximation is asymptotically exact as T grows large relative to the kernel's correlation length, and inherits boundary error at finite T. For our criteria the diagonal piece suffices because the criteria depend only on the per-basis-function eigenvalue λ_m and the data projection.

Inference under Gaussian likelihood A Gaussian variational posterior $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, S)$ on the inducing variables $\mathbf{u} = (u_j)_{j \in \mathcal{M}}$, the per-feature coefficients with prior variance $\lambda_j(\theta)$, is fit by maximising the standard sparse-GP ELBO. For Gaussian likelihood the optimum admits a closed form (the collapsed Titsias bound), and the bound itself becomes the marginal-likelihood objective that we maximise over θ by L-BFGS-B [25, 7]. The closed-form gradients and predictive expressions follow Hensman et al. directly.

Truncation baseline The conventional VFF library default fixes a per-axis frequency count $M_\star$ and keeps frequencies $j_d \in \{0, 1, \ldots, M_\star - 1\}$ on every axis. Each non-DC frequency contributes both a cosine and a sine basis function while $j_d = 0$ contributes only a cosine, so the total count is $M = D(2M_\star - 1)$. For each dataset we sweep $M_\star$ such that $M \in [16, 256]$ and thin to five log-spaced points (e.g. on *combined cycle power plant* with $D = 4$ this lands at $M \in \{20, 36, 60, 124, 252\}$, and on *kin8nm* with $D = 8$ at $M \in \{24, 40, 56, 120, 248\}$). This is the truncation curve reported in Figure 3.

E Variational inducing spherical harmonics (VISH)

For VISH [5, 6], the candidate dictionary is organised by spherical-harmonic degree after the Euclidean inputs have been lifted to the sphere. The details below give that construction, the order-1

arc-cosine kernel [3] used throughout, the closed-form zero pattern responsible for the cumulative-shell jump in Figure 4, and the two truncation baselines used in Section 6.

Input lifting and spherical harmonics Given Euclidean inputs $x_n \in \mathbb{R}^p$, define

$$z_n \;=\; \frac{(x_n, 1)}{\|(x_n, 1)\|} \;\in\; S^{D-1}, \qquad D = p + 1.$$

All harmonic features are evaluated at z_n, not at the raw Euclidean input. Spherical harmonics $\phi_{\ell,k} : S^{D-1} \to \mathbb{R}$ are indexed by degree $\ell = 0, 1, 2, \ldots$ and orientation $k = 1, \ldots, N(D, \ell)$, with multiplicity

$$N(D, \ell) \;=\; \frac{(2\ell + D - 2)\,(\ell + D - 3)!}{\ell!\,(D - 2)!}, \qquad D \geq 3,$$

and are orthonormal under the normalised surface measure $d\mu(z) = d\omega(z)/\Omega_{D-1}$.

Funk–Hecke and degree-only eigenvalues A kernel k_θ on S^{D-1} is zonal if $k_\theta(z, z') = \kappa_\theta(z^\top z')$, i.e. depends only on the angle between inputs. For a zonal kernel, the Funk–Hecke theorem says that spherical harmonics are eigenfunctions of the kernel integral operator with eigenvalues that depend only on the degree:

$$\lambda_{\ell,k}(\theta) \;=\; \lambda_\ell(\theta) \quad \text{for all } k = 1, \ldots, N(D, \ell).$$

This gives a shell decomposition indexed by degree ℓ, where all $N(D, \ell)$ orientations within a shell share the same prior variance.

The order-1 arc-cosine kernel We use the order-1 arc-cosine kernel [3] throughout, with $\lambda_\ell(\theta)$ obtained from the gpfy reference implementation [6]. A closed-form property of this kernel is that its Funk–Hecke expansion has $\lambda_\ell = 0$ for all odd $\ell \geq 3$, independent of the sphere dimension D [1, Appendix D.2]. Only the shells $\ell \in \{0, 1\} \cup \{2, 4, 6, \ldots\}$ carry positive prior variance. In the dimensions used by the lifted UCI datasets ($D \in \{5, 6, 7, 9, 12\}$), the same pattern is visible numerically: the magnitudes of the nonzero shells decay smoothly with ℓ, while the odd shells $\ell \geq 3$ sit at machine precision. The pattern is a property of the kernel function, not the sphere geometry, and it has a sharp practical consequence for the cumulative-shell truncation baseline below.

Inference under Gaussian likelihood Inducing variables are RKHS projections, $u_{\ell,k} = \langle f, \phi_{\ell,k} \rangle_{\mathcal{H}}$. The prior K_{uu} is diagonal in the spherical-harmonic basis with entries $1/\lambda_\ell$ in the RKHS coordinate. The Mercer coordinate $c_{\ell,k} = \lambda_\ell u_{\ell,k}$ has prior variance λ_ℓ, which is what the criteria use. A Gaussian variational posterior $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, S)$ is fit by maximising the standard sparse-GP ELBO, which for Gaussian likelihood admits a closed form. We maximise the resulting collapsed bound over θ by L-BFGS-B [5, 25]. The diagonal K_{uu} structure is preserved under arbitrary subset selection.

Two truncation baselines. The VISH comparisons use the two practitioner-default truncation rules from the literature.

The cumulative-shell rule of Dutordoir et al. [5] fills harmonic shells in degree order: pick a cutoff degree $L^\star$ and keep every harmonic at $\ell \leq L^\star$, for a total of $M = \sum_{\ell=0}^{L^\star} N(D, \ell)$ basis functions. This is the direct VISH analogue of the per-axis cube rule used for HSGP. On the order-1 arc-cosine kernel, however, this rule can select zero-variance shells: whenever $L^\star$ is an odd degree $\ell \geq 3$, every harmonic at that degree has zero prior variance and contributes no information to the fit. This is the cause of the upward jump in the black truncation curve in Figure 4.

The phase-truncation rule of Eleftheriadis et al. [6] keeps $\min(m^\star, N(D, \ell))$ orthogonalised phase features per shell up to a larger cutoff degree, where the phases are parameterised by learnable phase vectors and computed via the addition theorem as Gegenbauer evaluations $C_\ell^{(\alpha)}(z^\top v_{\ell,m})$ with $\alpha = (D - 2)/2$. This rule has a single integer knob $m^\star$ on top of the cutoff, and we choose $m^\star$ so that the resulting basis-function count is closest to the target M. It is data-blind (the phases are optimised at fitting time, not at selection time) and inherits the same odd-shell trap as cumulative-shell truncation.

F Toy figure details

The two illustrative figures in the main text serve a narrower purpose than the UCI sweeps: they make visible what a ranking rule is doing before any benchmark comparison is considered. We give the data-generation and fitting details for Figure 1 (the four-panel grid in Section 1) and Figure 5 (the one-dimensional bandpass example in Section 3) so that the examples can be read as concrete constructions rather than as additional empirical claims.

Figure 1 (motivational grid, two-dimensional HSGP) We use the AT and V features (axes 0 and 1) of the UCI *combined cycle power plant* regression [8] to build a $D = 2$ HSGP problem with $M_{\text{per-dim}} = 5$, giving a $5 \times 5 = 25$ candidate basis on a box $[-L_d, L_d]^2$ with $L_d = 1.2 \cdot \max_i |x_{id}|$ per axis. The kernel is Matérn-$5/2$ with ARD lengthscales. For each of the four selection strategies (rectangular truncation $M_d = (5, 2)$, the (no data) criterion I_j, the (no prior) criterion P_j, the (in-between) criterion $\tilde{H}_j$), we select the top-$M = 10$ candidates at the unit starting point θ_0 and fit the kernel hyperparameters by L-BFGS-B on the closed-form marginal likelihood from the same starting point, reporting test NLL on the held-out split (seed 0).

Figure 5 (one-dimensional bandpass example) We sample N noisy observations of a synthetic bandpass-sinc signal on $t \in [0, T]$ at sampling frequency $f_s = 512$ ($T = 1$, $N = 512$ regularly-spaced points), with bandpass cutoffs $\omega_0 = 30$ and $\omega_1 = 100$ and additive Gaussian noise of standard deviation $\sigma_{\text{noise}} = 0.3$. The HSGP basis uses $M_{\text{cand}} = 512$ Dirichlet sine eigenfunctions on a box of length $1.2 \cdot T$. The score-spectrum panels show I_j and $\tilde{H}_j$ at $M = 10$ alongside the corresponding HSGP fits.

G Experimental details for UCI

Datasets and splits Six UCI regression benchmarks [8]: airfoil self-noise ($N = 1503$, $D = 5$), concrete compressive strength ($N = 1030$, $D = 8$), energy efficiency ($N = 768$, $D = 8$), kin8nm ($N = 8192$, $D = 8$), combined cycle power plant ($N = 9568$, $D = 4$), and yacht hydrodynamics ($N = 308$, $D = 6$). Each cell uses 10 random 90:10 train/test splits with seeds $0, \ldots, 9$. Inputs and targets are standardised to zero mean and unit variance using training-set statistics.

Optimisation protocol (HSGP, VFF, VISH) For HSGP, VFF, and VISH, each cell is fit with a single L-BFGS-B run over $(\log \ell, \log \sigma^2_{\text{sig}}, \log \sigma^2_{\text{noise}})$ from the unit starting point $\theta_0 = (\log \ell_d = 0, \log \sigma^2_{\text{sig}} = 0, \log \sigma^2_{\text{noise}} = \log 0.1)$. The same starting point is used for selection and for the subsequent kernel-hyperparameter fit. We do not multi-start. Variability across the 10 splits is reported in every per-family figure as the median test NLL with shaded interquartile bands, robust to occasional optimisation outliers.

HSGP setup For HSGP, we use a Matérn-$5/2$ ARD kernel on a box $[-L_d, L_d]^D$ per axis with $L_d = 1.2 \cdot \max_i |x_{id}|$. The candidate dictionary is $\{1, \ldots, M_{\text{per-dim}}\}^D$, with $M_{\text{per-dim}}$ chosen so that the total candidate count satisfies $J \leq 8000$. The fit objective is the closed-form marginal likelihood of the truncated linear-Gaussian model described in Appendix C. The non-uniform truncation baseline is constructed as documented in Appendix C.

VFF setup For VFF, we use the per-axis Matérn-$5/2$ K_{uu} blocks of Hensman et al. [7]. The candidate dictionary is $\{(d, j_d) : 1 \leq d \leq D,\ 0 \leq j_d < M_{\text{per-dim}}\}$. It is treated as a flat set and ranked globally. The fit objective is the closed-form collapsed Titsias bound described in Appendix D, and the truncation baseline is the one documented there.

VISH setup For VISH, inputs are projected to S^D via $z_n = (x_n, 1)/\|(x_n, 1)\|$ [5]. Spherical harmonics and the order-1 arc-cosine Funk–Hecke spectrum are taken from the gpfy library (Apache-2.0) [6, 3], with $\ell_{\max}$ chosen per D to fit gpfy's pre-computed fundamental-set tables. The fit objective is the collapsed Titsias bound on the gpfy spherical-harmonic basis described in Appendix E. Cumulative-shell truncation cuts at the largest degree $L^\star$ satisfying $\sum_{\ell \leq L^\star} N(D, \ell) \leq M$. The Eleftheriadis phase-truncation baseline uses gpfy's `phase_truncation` parameter, choosing the per-shell phase count $m^\star$ so the total basis-function count is closest to the target M.

Compute resources The paper experiments, meaning the runs that produced the JSON files used for the reported figures, were run on a single workstation with an AMD Ryzen 7 3700X CPU (8 cores, 16 hardware threads), 64 GB RAM, and an NVIDIA RTX 2080 SUPER GPU (8 GB). Summing the seven UCI sweep modules used in the paper gives approximately 3.9 hours for the six reported datasets, with individual L-BFGS-B fits typically completing in seconds.

H Test RMSE on UCI for Experiments I, II, III

Figures H.1, H.2, and H.3 repeat the UCI comparison with median test root mean square error (RMSE) versus budget M across the same 10 train/test splits, with interquartile bands. The pattern across the three families is the same as the test-NLL pattern in the main text. On HSGP the data-aware criteria $\tilde{H}_j$ and P_j improve on I_j on most cells. On VFF the three criteria are roughly tied and the per-axis truncation baseline tracks I_j closely. On VISH the (no data) I_j matches or exceeds the data-aware criteria, and the cumulative-shell truncation curve shows the same odd-shell jumps the test-NLL curve does. We use NLL as the headline metric in the body for two reasons: it captures the predictive distribution rather than only the point prediction, and it is the standard test metric for sparse GP regression in this literature.

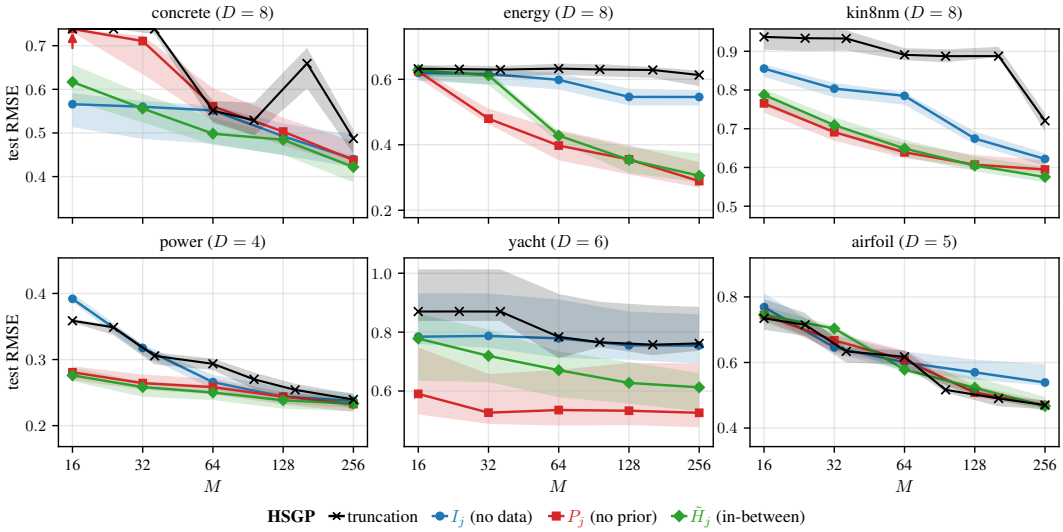

Figure H.1: Experiment I: HSGP test RMSE on the same six UCI benchmarks and budgets as Figure 2.

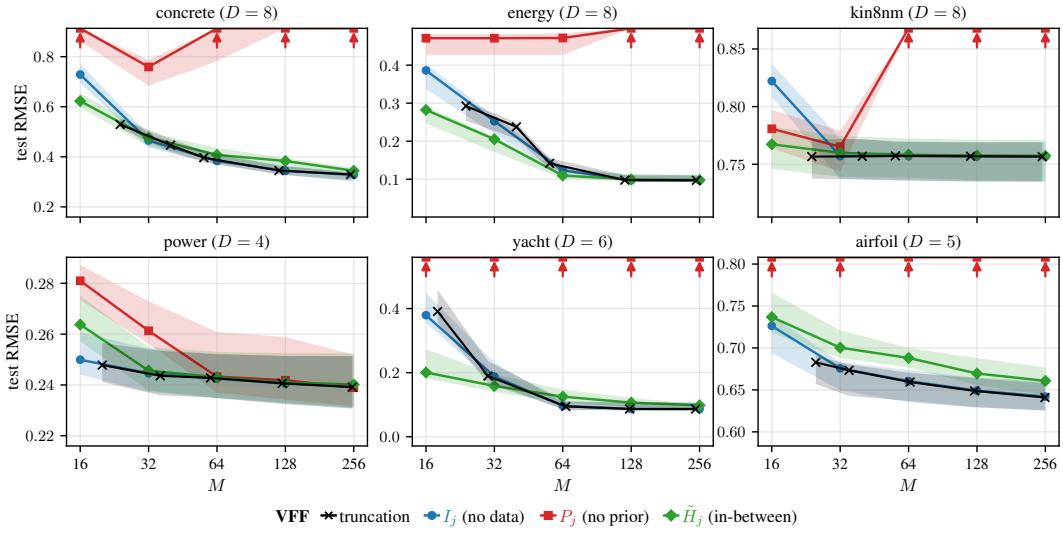

Figure H.2: Experiment II: VFF test RMSE on the same six UCI benchmarks and budgets as Figure 3.

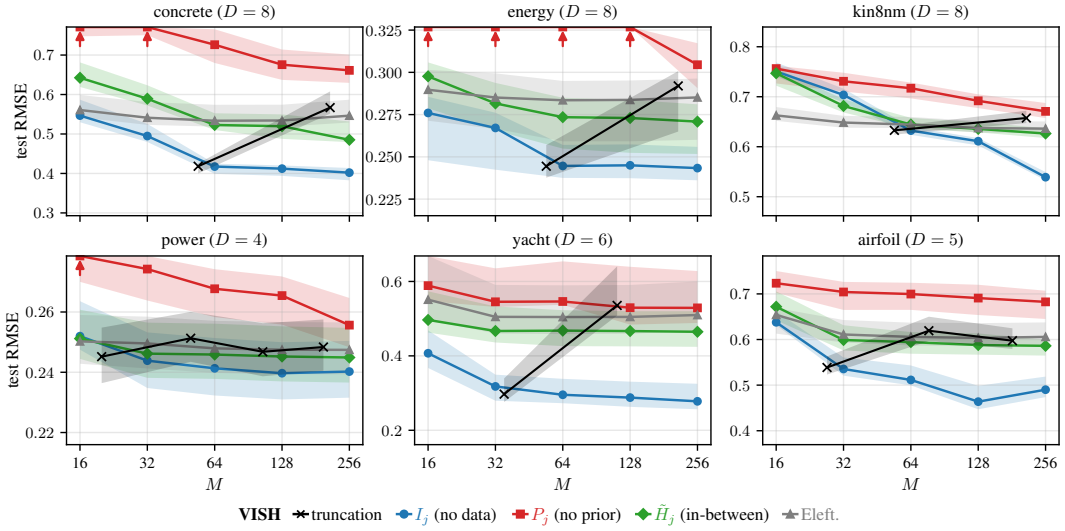

Figure H.3: Experiment III: VISH test RMSE on the same six UCI benchmarks and budgets as Figure 4.