

Information-Based Calibration of Uncertainty Quantification in Product-of-Experts Gaussian Process Models

YEAN HOON ONG*, University College London, UK

PAOLO BARUCCA, University College London, UK

WEI PAN, The University of Manchester, UK

JUN WANG, University College London, UK

Background: In Gaussian process (GP) regression, the use of a single global GP (GP-glo) incurs cubic computational cost, which limits scalability to large datasets. Product-of-experts GP models (GP-pro), which combine a collection of local GP models that collaboratively capture global correlations, provide a prominent approach to alleviating this computational burden. However, GP-pro models often produce overestimated posterior variances due to training local experts on disjoint subsets of the data.

Objectives: This study aims to analyse the overestimation of posterior variances in GP-pro models and to address this issue using an information-based method.

Methods: We propose GP-pro-c, a product-of-experts Gaussian process model that incorporates an information-based method to calibrate the overestimated posterior variances. The method uses the monotonicity and submodularity properties of information gain in GP to define a calibration ratio that appropriately reduces the posterior variance of individual local GP models.

Results: The performance of GP-pro-c was evaluated using three metrics: negative log-likelihood (NLL), root mean squared error (RMSE), and expected normalised calibration error (ENCE) of the uncertainty estimates. Empirical experiments on four synthetic functions and six regression datasets show that the proposed calibration method enables the GP-pro-c model to achieve average reductions of 2.3% and 12.0% in NLL and ENCE, respectively, compared to the uncalibrated GP-pro model.

Conclusions: The results demonstrate that GP-pro-c effectively mitigates the overestimation of posterior variances in product-of-experts Gaussian process models while maintaining predictive accuracy and reducing computational complexity. These findings suggest that the proposed information-based calibration method is a promising approach for improving uncertainty estimation in scalable GP models. We expect the GP-pro-c to serve as a useful surrogate model in Bayesian optimisation settings involving high-dimensional and large-scale data.

JAIR Associate Editor: Tongliang Liu

JAIR Reference Format:

Yean Hoon Ong, Paolo Barucca, Wei Pan, and Jun Wang. 2026. Information-Based Calibration of Uncertainty Quantification in Product-of-Experts Gaussian Process Models. *Journal of Artificial Intelligence Research* 86, Article 51 (August 2026), 30 pages. DOI: [10.1613/jair.1.20374](https://doi.org/10.1613/jair.1.20374)

*Corresponding Author.

Authors' Contact Information: Yean Hoon Ong, ORCID: [0000-0002-8452-0999](https://orcid.org/0000-0002-8452-0999), yeen-hoon.ong@ucl.ac.uk, University College London, London, UK; Paolo Barucca, p.barucca@ucl.ac.uk, ORCID: [0000-0003-4588-667X](https://orcid.org/0000-0003-4588-667X), University College London, London, UK; Wei Pan, wei.pan@manchester.ac.uk, ORCID: [0000-0003-1121-9879](https://orcid.org/0000-0003-1121-9879), The University of Manchester, Manchester, UK; Jun Wang, jun.l.wang@ucl.ac.uk, ORCID: [0000-0002-4021-4228](https://orcid.org/0000-0002-4021-4228), University College London, London, UK.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2026 Copyright held by the owner/author(s).

DOI: [10.1613/jair.1.20374](https://doi.org/10.1613/jair.1.20374)

1 Introduction

Gaussian process (GP) models (Rasmussen and Williams 2006) are nonparametric methods widely used for nonlinear regression. They provide both posterior predictions and uncertainty quantification, which support principled decision-making. A single global GP model is defined by a mean function and a covariance function. Given n training data points, model fitting and inference require inversion of the covariance matrix, resulting in cubic computational cost $O(n^3)$ and quadratic memory requirements $O(n^2)$. These limitations hinder the application of global GP models to large-scale problems.

A variety of approaches have been proposed to address the scalability limitations of GP models. These can be broadly categorised into three classes: sparse inducing-point methods, expert-based methods, and dimensionality-reduction methods. Sparse inducing-point approaches (H. Liu, Ong, et al. 2020; Nguyen and Bonilla 2014) reduce computational cost by summarising the dataset with a smaller set of representative points. While effective for moderate input dimensionality, they may struggle to capture highly localised structures or scale to very large datasets. Expert-based approaches (Cao 2018; Deisenroth and Ng 2015; H. Liu, Ong, et al. 2020; Tresp 2000) partition the input space and aggregate predictions from multiple local models. Dimensionality-reduction approaches, such as the principal component-based method of Higdon et al. (2008), address scalability from a different perspective by projecting high-dimensional outputs into a lower-dimensional subspace. This reduces the effective output dimensionality and enables efficient emulation of complex computer models. Unlike expert-based approaches, which primarily address scalability with respect to the number of observations and input structure, principal component-based methods focus on output compression; the two approaches are therefore complementary.

In this work, we focus on expert-based approaches, specifically the product-of-experts GP model, which combines a collection of local GP models (Cao 2018; Cao and Fleet 2014, 2015; Deisenroth and Ng 2015; H. Liu, Cai, et al. 2019; H. Liu, Ong, et al. 2020; Tresp 2000). The key idea is to partition the training data into M subsets, each containing $\tilde{n}$ points with $\tilde{n} \ll n$, and to train a GP model on each subset. At test time, predictions are aggregated by multiplying the outputs of the local experts. This reduces the training complexity to $O(M\tilde{n}^3)$ and the memory requirement to $O(M(\tilde{n}^2 + \tilde{n}d))$ (H. Liu, Ong, et al. 2020), where M is the number of local GP models, d denotes the input dimensionality, and assuming that each local GP uses the same number of $\tilde{n}$ training data points.

Three notable variants of the product-of-experts framework are the Bayesian committee machine (BCM) (Tresp 2000), the generalised product of experts (gPoE) (Cao 2018; Cao and Fleet 2014, 2015), and the robust Bayesian committee machine (rBCM) (Deisenroth and Ng 2015). The BCM and rBCM assume a shared GP prior across all experts and require identical hyperparameters, which can limit flexibility in modelling locally varying patterns (H. Liu, Cai, et al. 2019). Compared to gPoE, rBCM offers stronger Bayesian consistency and improved behaviour in data-sparse regions due to its principled prior correction. In contrast, gPoE provides several practical advantages: (i) it captures local patterns by allowing each expert to learn distinct hyperparameters; (ii) it enables flexible weighting of individual experts; (iii) it captures long-range correlations; and (iv) it reduces computational cost (Cao and Fleet 2014, 2015; Deisenroth and Ng 2015; H. Liu, Ong, et al. 2020). We note that while gPoE can adapt to data-induced heterogeneity across experts, this is not equivalent to modelling true nonstationary kernels. For the above-mentioned practical reasons, we focus on the gPoE model. For brevity, we use GP-pro to refer to the generalised product-of-experts model unless otherwise stated.

Despite its advantages, GP-pro models tend to overestimate predictive uncertainty (variance). This occurs because each local GP is trained on only a subset of the data: while this reduces computational cost, it also limits the information available to each expert, causing each local predictive variance to be larger than it would be if the GP were trained on all data. Since GP predictive variance decreases as more training data are incorporated (Desautels et al. 2014), aggregating predictions from multiple local experts without correction propagates this overestimation to the final prediction (Deisenroth and Ng 2015; H. Liu, Ong, et al. 2020; Szabó and Zanten 2019).

Such miscalibration can negatively affect downstream tasks, including Bayesian optimisation, active learning, and other uncertainty-sensitive applications.

Researchers have proposed the use of a scaling temperature to calibrate the predictive uncertainties in GP-pro models (Cohen et al. 2020). In this approach, the predictive variances of all local GPs are multiplied by a constant factor to adjust their magnitude and better reflect the true variability of the function. However, the method requires the user to pre-define this temperature constant, and if it is set too high or too low, the model can become overly uncertain or overconfident, meaning that the predicted variances still do not accurately match the actual uncertainty.

In this work, we address variance overestimation in GP-pro models by proposing an information-based calibration method. The method adjusts the posterior variance of each local expert at test time using quantities already available in standard GP-pro predictions. It leverages the monotonicity and submodularity of information gain to determine an appropriate calibration factor. This results in improved uncertainty estimates while preserving computational efficiency. Empirical results demonstrate that the proposed approach yields well-calibrated uncertainty without compromising predictive accuracy.

This study makes four main contributions:

- An analysis of variance overestimation in GP-pro models.
- An information-based calibration method for uncertainty quantification in GP-pro, resulting in the GP-pro-c model.
- A comprehensive comparison of GP-pro and GP-pro-c across multiple data assignment methods and weighting schemes.
- An empirical evaluation of GP-pro-c against GP-pro, temperature-scaled GP-pro, and a global GP model in terms of predictive performance and computational cost.

The remainder of this paper is organised as follows. Section 2 introduces background on global GP models, GP-pro, and relevant information-theoretic concepts. Section 3 reviews related work. Section 4 analyses variance overestimation in GP-pro. Section 5 presents the proposed information-based calibration method. Section 6 reports experimental results, and Section 7 concludes the paper.

2 Preliminaries

A Gaussian process (GP) is a nonparametric probabilistic model that defines a distribution over functions via a collection of random variables. The reader is referred to Rasmussen and Williams (2006) for a comprehensive introduction. A GP is specified by a mean function and a covariance function, and is written as:

$$f(\mathbf{x}) \sim GP(\tilde{\mu}(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

where $f(\mathbf{x})$ is a real-valued stochastic process, and $\tilde{\mu}(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ denote the mean and covariance functions, respectively, defined as:

$$\begin{aligned}\tilde{\mu}(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - \tilde{\mu}(\mathbf{x}))(f(\mathbf{x}') - \tilde{\mu}(\mathbf{x}'))]\end{aligned}$$

The hyperparameters of the prior mean function and the covariance function, denoted by θ , are learned from data. Given a set of n training data points and their corresponding observations $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i \leq n}$, where $\mathbf{x}_i$ denotes an input vector with dimension d and y_i denotes a scalar output or target. The observed value y_i is assumed to differ from the function values $f(\mathbf{x}_i)$ by additive noise, and this noise is assumed to follow an independent, identically distributed Gaussian distribution with zero mean and variance σ^2 , i.e., $y_i = f(\mathbf{x}_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The hyperparameters θ are estimated by maximising the log marginal likelihood:

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2}(\mathbf{y} - \tilde{\boldsymbol{\mu}})^\top (\mathbf{K} + \tilde{\sigma}^2 \mathbf{I})^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}) - \frac{1}{2} \log |\mathbf{K} + \tilde{\sigma}^2 \mathbf{I}| - \frac{n}{2} \log 2\pi \quad (1)$$

where $\tilde{\boldsymbol{\mu}}$ is the vector of mean function evaluations, and $\mathbf{K}$ is the covariance matrix with entries $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

For a test input $\mathbf{x}'$, the predictive distribution of $y' = f(\mathbf{x}') + \epsilon'$ is Gaussian:

$$p(y' | \mathcal{D}, \mathbf{x}') = \mathcal{N}(\mu(\mathbf{x}'), \sigma^2(\mathbf{x}') + \tilde{\sigma}^2)$$

where the posterior mean and variance are given by:

$$\mu(\mathbf{x}') = \tilde{\boldsymbol{\mu}}(\mathbf{x}') + \mathbf{k}(\mathbf{x}')^\top (\mathbf{K} + \tilde{\sigma}^2 \mathbf{I})^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}) \quad (2)$$

$$\sigma^2(\mathbf{x}') = k(\mathbf{x}', \mathbf{x}') - \mathbf{k}(\mathbf{x}')^\top (\mathbf{K} + \tilde{\sigma}^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}') \quad (3)$$

and $\mathbf{k}(\mathbf{x}') = [k(\mathbf{x}', \mathbf{x}_1), \dots, k(\mathbf{x}', \mathbf{x}_n)]^\top$ denotes the covariance vector between the test input and the training inputs.

2.1 Product-of-Experts Gaussian Process Models (GP-pro)

The GP-pro model consists of a collection of $M > 1$ local GP models, where each local GP models a distinct region of the input domain and performs independent predictions and uncertainty estimation. The overall predictive distribution is obtained by combining the outputs of the individual local GPs through a product-of-experts formulation. The two essential components of a GP-pro model are (i) the data assignment method used to allocate training data to each local GP and (ii) the weighting scheme used to aggregate the posterior predictions and uncertainties provided by each expert.

Assigning data and training the GP-pro model. Let N_{GP} denote the maximum number of data points assigned to each local GP model. Given a training dataset $\mathcal{D}$, the number of local GP models is determined as $M = \text{int}(|\mathcal{D}|/N_{GP})$. Each local $GP^{(m)}$ is assigned a subset $\mathcal{D}^{(m)} \subset \mathcal{D}$. Common data assignment methods include random partitioning, k -means clustering, and balltree methods (Cao 2018; Cao and Fleet 2014; Deisenroth and Ng 2015; H. Liu, Cai, et al. 2019).

With random assignment, each local $GP^{(m)}$ is allocated the equal number of training data points, i.e., $|\mathcal{D}^{(m)}| = |\mathcal{D}|/M$. Data points within each $\mathcal{D}^{(m)}$ are sampled uniformly (with or without replacement) from $\mathcal{D}$.

With k -means assignment, the training inputs $\{\mathbf{x}_i\}_{i \leq n}$ are partitioned into M disjoint clusters using the k -means algorithm (Arthur and Vassilvitskii 2007). Each local GP is then trained on one cluster, resulting in potentially uneven subset sizes.

With balltree assignment, a balltree (Omohundro 1989) is constructed such that the training data points are partitioned into a nested set of balls by recursively splitting the points into two disjoint sets based on a distance metric. In this study, each leaf node contains at least $|\mathcal{D}|/M$ data points. Traversing from the leaf nodes to the parent nodes, M disjoint subsets of data are constructed by randomly choosing $|\mathcal{D}^{(m)}|$ data points without replacement from individual nodes. Each subset of data $\mathcal{D}^{(m)}$ is used to fit a local $GP^{(m)}$ model.

The random assignment method is computationally efficient, incurring only linear-time complexity in the number of data points. The k -means and balltree methods introduce additional overhead due to clustering or spatial indexing, but this cost is negligible relative to the cubic $\mathcal{O}(n^3)$ training cost of a global GP. In this study, we use the k -means implementation from SciPy¹ and the balltree implementation from scikit-learn². Empirical runtime results later in the paper confirm that the overhead of these assignment methods is minor compared to GP training and inference.

¹<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.vq.kmeans2.html>

²<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.html>

Each local GP model $GP^{(m)}$ is trained on the respective $\mathcal{D}^{(m)}$ using a constant prior mean function and a Matérn-5/2 covariance function. The hyperparameters of $GP^{(m)}$, denoted by $\theta^{(m)}$, are estimated by maximising the log marginal likelihood. The marginal likelihood for each local GP expert has the same functional form as the global GP marginal likelihood in Equation (1), but is evaluated independently for each local dataset $\mathcal{D}^{(m)}$ with expert-specific hyperparameters $\theta^{(m)}$:

$$\begin{aligned} \log p(\mathbf{y}^{(m)} | \mathbf{X}^{(m)}, \theta^{(m)}) = & -\frac{1}{2} \left(\mathbf{y}^{(m)} - \tilde{\boldsymbol{\mu}}^{(m)} \right)^\top \left(\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I} \right)^{-1} \left(\mathbf{y}^{(m)} - \tilde{\boldsymbol{\mu}}^{(m)} \right) \\ & - \frac{1}{2} \log \left| \mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I} \right| - \frac{|\mathcal{D}^{(m)}|}{2} \log 2\pi \end{aligned} \quad (4)$$

where $\tilde{\boldsymbol{\mu}}^{(m)}$ is a vector of mean function values for $\mathbf{X}^{(m)}$, $\ddot{\sigma}_m^2$ is the noise variance, and $\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I}$ is the covariance matrix for the noisy target $\mathbf{y}^{(m)}$.

Aggregating posterior predictions. At test time, at candidate $\mathbf{x}'$, each local $GP^{(m)}$ computes its posterior predictive mean $\mu_m(\mathbf{x}')$ and variance $\sigma_m^2(\mathbf{x}')$ independently. These predictive expressions have the same functional form as the global GP predictive mean and variance in Equations (2) and (3), but is evaluated independently for each local dataset $\mathcal{D}^{(m)}$ with expert-specific hyperparameters $\theta^{(m)}$:

$$\mu_m(\mathbf{x}') = \tilde{\mu}^{(m)}(\mathbf{x}') + \mathbf{k}^{(m)}(\mathbf{x}')^\top \left(\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I} \right)^{-1} \left(\mathbf{y}^{(m)} - \tilde{\boldsymbol{\mu}}^{(m)} \right) \quad (5)$$

$$\sigma_m^2(\mathbf{x}') = k^{(m)}(\mathbf{x}', \mathbf{x}') - \mathbf{k}^{(m)}(\mathbf{x}')^\top \left(\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I} \right)^{-1} \mathbf{k}^{(m)}(\mathbf{x}') \quad (6)$$

where $\tilde{\mu}^{(m)}$ is the mean function, $k^{(m)}$ is the covariance function, $\mathbf{k}^{(m)}$ is the vector of covariance terms between $\mathbf{x}'$ and $\mathbf{X}^{(m)}$, and $\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I}$ is the covariance matrix.

Following [Cao and Fleet \(2014\)](#), the aggregated predictive distribution is defined as a product of expert predictions:

$$p(y' | \mathbf{x}', \mathcal{D}) = \prod_{m=1}^M p^{w_m(\mathbf{x}')} (y' | \mathbf{x}', \mathcal{D}^{(m)})$$

This yields the aggregated predictive mean and variance:

$$\mu(\mathbf{x}') = \sigma^2(\mathbf{x}') \sum_{m=1}^M \left(w_m(\mathbf{x}') \mu_m(\mathbf{x}') \frac{1}{\sigma_m^2(\mathbf{x}')} \right) \quad (7)$$

$$\sigma^2(\mathbf{x}') = \left(\sum_{m=1}^M \left(w_m(\mathbf{x}') \frac{1}{\sigma_m^2(\mathbf{x}')} \right) \right)^{-1} \quad (8)$$

The weight $w_m(\mathbf{x}') \in \mathbb{R}^+$ reflects the reliability of the m -th expert at $\mathbf{x}'$. The higher the weight $w_m(\mathbf{x}')$, the more reliable the m -th expert will be. The weighting factor at each $\mathbf{x}'$ is normalised such that $\sum_{m'=1}^M w_{m'}(\mathbf{x}') = 1$. We consider three weighting schemes: entropy-based, variance-based, and uniform weighting.

The entropy-weighted scheme (ENT) ([Cao and Fleet 2014](#)) measures reliability as the reduction in differential entropy between the prior and posterior. For the Gaussian distribution, the differential entropy of local $GP^{(m)}$ at candidate $\mathbf{x}'$ can be computed in closed form as $H_m(\mathbf{x}') = \frac{1}{2} (1 + \log 2\pi \sigma_m^2(\mathbf{x}'))$. With that, the quantity $\Delta H_m(\mathbf{x}')$

is given as:

$$\begin{aligned}\Delta H_m(\mathbf{x}') &= \tilde{H}_m(\mathbf{x}') - H_m(\mathbf{x}') \\ &= \frac{1}{2} (1 + \log 2\pi \tilde{\sigma}_m^2(\mathbf{x}')) - \frac{1}{2} (1 + \log 2\pi \sigma_m^2(\mathbf{x}')) \\ &= \frac{1}{2} (\log \tilde{\sigma}_m^2(\mathbf{x}') - \log \sigma_m^2(\mathbf{x}'))\end{aligned}$$

where $\tilde{\sigma}_m^2(\mathbf{x}') = k^{(m)}(\mathbf{x}', \mathbf{x}')$ is the prior variance, and $\sigma_m^2(\mathbf{x}')$ is the posterior predictive variance of local $GP^{(m)}$ at candidate $\mathbf{x}'$. Using the quantity $\Delta H_m(\mathbf{x}')$, the normalised weight of a local $GP^{(m)}$ at candidate $\mathbf{x}'$ is:

$$w_m^{\text{ENT}}(\mathbf{x}') = \frac{\Delta H_m(\mathbf{x}')}{\sum_{m'=1}^M \Delta H_{m'}(\mathbf{x}')} \quad (9)$$

A higher difference in entropy indicates a higher reliability of a local $GP^{(m)}$, and therefore the posterior prediction provided by this local GP model should contribute more to the aggregated posterior prediction.

With the variance-weighted scheme (VAR), the reliability of local $GP^{(m)}$ is measured as the difference between the prior and posterior variances at candidate $\mathbf{x}'$. The higher the difference in variance, the more reliable the local $GP^{(m)}$. Therefore, at candidate $\mathbf{x}'$, the normalised weighting factor w^{VAR} of a local $GP^{(m)}$ is:

$$w_m^{\text{VAR}}(\mathbf{x}') = \frac{\tilde{\sigma}_m^2(\mathbf{x}') - \sigma_m^2(\mathbf{x}')}{\sum_{m'=1}^M (\tilde{\sigma}_{m'}^2(\mathbf{x}') - \sigma_{m'}^2(\mathbf{x}'))} \quad (10)$$

where $\tilde{\sigma}_m^2(\mathbf{x}')$ is the prior variance and $\sigma_m^2(\mathbf{x}')$ is the posterior predictive variance of the local $GP^{(m)}$ at the candidate $\mathbf{x}'$.

The variance weighting is different from the entropy weighting in that the difference in entropy is unitless, while the difference in variance carries the unit of variance (Cao and Fleet 2014). Moreover, because the variance-weighted scheme depends directly on the magnitude of the posterior variance reduction, it tends to more strongly downweight local $GP^{(m)}$ models that remain highly uncertain at the candidate $\mathbf{x}'$. In contrast, entropy-based weighting induces a smoother penalisation, as entropy grows logarithmically with variance. Neither weighting scheme is universally preferable; variance weighting emphasises sensitivity to uncertainty magnitude, while entropy weighting yields smoother aggregation. Our proposed information-based calibration method is orthogonal to this choice and can be applied in conjunction with either weighting scheme.

For comparison, the uniform-weighted scheme (UNI) assigns equal weight to all experts:

$$w_m^{\text{UNI}}(\mathbf{x}') = \frac{1}{M} \quad (11)$$

Figure 1 (right) illustrates an example GP-pro model with two local experts on the one-dimensional Ackley function, using balltree assignment and entropy-weighted aggregation.

2.2 Information Gain for Covariance Functions

In a GP model, the covariance function (kernel) is a fundamental component that encodes prior assumptions about the objective function. In this section, we introduce key information-theoretic quantities that form the basis of our proposed calibration method for addressing variance overestimation in GP-pro models.

In the framework of Bayesian experimental design (Chaloner and Verdinelli 1995), the informativeness of a set of sampling points $\mathcal{A} \subset \mathcal{X}$ about the latent function f is quantified by the information gain (Cover and Thomas 2005). Information gain measures the reduction in uncertainty about f induced by observing data. Specifically, for a set $\mathcal{A}$, we observe noisy outputs $\mathbf{y}_{\mathcal{A}} = \mathbf{f}_{\mathcal{A}} + \epsilon_{\mathcal{A}}$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ denotes independent Gaussian noise. Observing $\mathbf{y}_{\mathcal{A}}$ reduces the uncertainty about the function values $\mathbf{f}$ over the entire domain $\mathcal{X}$.

The information gain associated with $\mathcal{A}$ is defined as the mutual information between $\mathbf{y}_{\mathcal{A}}$ and f :

$$I(\mathbf{y}_{\mathcal{A}}; f) = H(f) - H(f | \mathbf{y}_{\mathcal{A}}) = H(\mathbf{y}_{\mathcal{A}}) - H(\mathbf{y}_{\mathcal{A}} | f)$$

For Gaussian distributions, the differential entropy is given by $H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \log |2\pi e \Sigma|$, and the mutual information can be expressed in terms of a log-determinant:

$$\begin{aligned} I(\mathbf{y}_{\mathcal{A}}; f) &= I(\mathbf{y}_{\mathcal{A}}; f_{\mathcal{A}}) = H(\mathbf{y}_{\mathcal{A}}) - H(\mathbf{y}_{\mathcal{A}} | f) \\ &= \frac{1}{2} \log |I + \tilde{\sigma}^{-2} K_{\mathcal{A}}| \end{aligned} \quad (12)$$

where $K_{\mathcal{A}} = [k(\mathbf{x}, \mathbf{x}')]_{\mathbf{x}, \mathbf{x}' \in \mathcal{A}}$ is the covariance matrix of $f_{\mathcal{A}}$. For GP models, $\mathbf{y}_{\mathcal{A}}$ depends only on $f_{\mathcal{A}}$, which implies $H(\mathbf{y}_{\mathcal{A}} | f) = H(\mathbf{y}_{\mathcal{A}} | f_{\mathcal{A}})$ and hence $I(\mathbf{y}_{\mathcal{A}}; f) = I(\mathbf{y}_{\mathcal{A}}; f_{\mathcal{A}})$ (Desautels et al. 2014; Srinivas et al. 2012).

The information gain is both monotonic and submodular (Cover and Thomas 2005; Krause et al. 2008; Srinivas et al. 2012). Monotonicity implies that for $\mathcal{A} \subset \mathcal{A}'$, we have $I(\mathbf{y}_{\mathcal{A}}; f) \leq I(\mathbf{y}_{\mathcal{A}'}; f)$. Submodularity captures a diminishing returns property: adding a new data point to a larger set yields a smaller marginal increase in information gain than adding it to a smaller set.

Let γ_N denote the maximum information gain obtainable from any subset $\mathcal{A} \subset \mathcal{X}$ with $|\mathcal{A}| \leq N$, defined as $\gamma_N = \max_{\mathcal{A} \subset \mathcal{X}, |\mathcal{A}| \leq N} I(\mathbf{y}_{\mathcal{A}}; f)$. Computing γ_N exactly is NP-hard. However, a greedy algorithm provides an efficient approximation (Dorard 2012; Ko et al. 1995; Nemhauser et al. 1978; Srinivas et al. 2012). At iteration n , the algorithm selects $\mathbf{x}_n = \arg \max_{\mathbf{x} \in \mathcal{X}} I(\mathbf{y}_{\mathcal{A}_{n-1} \cup \{\mathbf{x}\}}; f)$, where $\mathcal{A}_{n-1} = \{\mathbf{x}_1, \dots, \mathbf{x}_{n-1}\}$. For GP models, this selection is equivalent to choosing the point with maximum posterior variance: $\mathbf{x}_n = \arg \max_{\mathbf{x} \in \mathcal{X}} \sigma_{n-1}^2(\mathbf{x})$.

Due to submodularity, Nemhauser et al. (1978) showed that the information gain achieved by the greedy selection, denoted $I^G(\mathbf{y}_{\mathcal{A}}; f)$, satisfies:

$$I^G(\mathbf{y}_{\mathcal{A}}; f) \geq \left(1 - \frac{1}{e}\right) \max_{\mathcal{A} \subset \mathcal{X}, |\mathcal{A}| \leq N} I(\mathbf{y}_{\mathcal{A}}; f) \quad (13)$$

Furthermore, Dorard (2012) showed that the greedy information gain can be expressed as a sum of marginal conditional gains:

$$I^G(\mathbf{y}_{\mathcal{A}}; f) = \frac{1}{2} \sum_{n=1}^N \log \left(1 + \frac{\sigma_{n-1}^2(\mathbf{x}_n)}{\tilde{\sigma}^2}\right) \quad (14)$$

where $\sigma_{n-1}^2(\mathbf{x}_n)$ is the posterior variance at $\mathbf{x}_n$ conditioned on $\{y_1, \dots, y_{n-1}\}$. A proof is provided in Appendix A.

The maximum information gain γ_N depends on both the kernel and the input space (Srinivas et al. 2012). Using the greedy bound in Equation (13), Srinivas et al. (2012) derived highly problem-specific bounds on γ_N , which are simple analytical expressions of N and the power of k (covariance function), for three most commonly used kernels. For a linear kernel, $\gamma_N = O(d \log N)$. For the squared exponential kernel, $\gamma_N = O((\log N)^{d+1})$. For the Matérn kernel with $\nu > 1$, $\gamma_N = O(N^{\frac{d(d+1)}{2\nu+d(d+1)}} (\log N)) + O(N^0)$.

3 Related Work

The GP-pro model depends on both a data assignment method for distributing data among local GPs and a weighting scheme for aggregating their posterior predictions. Cao and Fleet (2014) conducted a comparative study of random, k -means, and balltree assignment methods combined with an entropy-based weighting scheme, reporting that the balltree assignment achieved slightly better performance. Most subsequent work (Deisenroth and Ng 2015; H. Liu, Ong, et al. 2020; Z. Liu et al. 2016) adopts random assignment with entropy-based weighting, while Cohen et al. (2020) considered k -means assignment with variance-based weighting. To the best of our knowledge, no systematic comparison has been conducted across different combinations of assignment methods and weighting schemes. In this study, we perform a comprehensive evaluation of GP-pro variants (with and

without calibration) using three assignment methods (random, k -means, and balltree) and three weighting schemes (entropy, variance, and uniform), examining both predictive performance and computational overhead.

Uncertainty calibration in GP models has been studied in several contexts. In computer model calibration, discrepancy-based approaches introduce an explicit model–data discrepancy term to account for simulator bias (e.g., [Higdon et al. \(2008\)](#)). These methods are particularly effective for physical simulations with high-dimensional outputs, but they address structural model mismatch rather than miscalibration arising from approximate inference. In contrast, scalable GP models based on local or distributed approximations, such as product-of-experts formulations, exhibit miscalibration due to the aggregation of multiple local posteriors. While these models alleviate cubic computational costs, they often produce inaccurate uncertainty estimates. The present work addresses this form of miscalibration by proposing an information-based variance calibration method for GP-pro models.

To mitigate overestimated variances in GP-pro, [Cohen et al. \(2020\)](#) proposed a weight calibration approach based on a temperature-scaled softmax function:

$$\tilde{w}_m(\mathbf{x}') \propto \exp(-T\psi_m(\mathbf{x}')), \quad \sum_{m=1}^M \tilde{w}_m(\mathbf{x}') = 1$$

Here, T is an inverse temperature parameter that controls the concentration of weights, amplifying stronger experts while downweighting weaker ones. The functional $\psi_m(\mathbf{x}')$ describes the level of confidence of the m -th expert at the test point $\mathbf{x}'$, typically defined using the posterior variance or differential entropy. However, this approach calibrates the aggregation weights rather than directly correcting the overestimated variances of individual experts. Moreover, the temperature parameter must be specified a priori, and poor choices can lead to undesirable behaviour: large T values overly concentrate weight on a small subset of experts, while small T values produce overly diffuse weights. Both cases can degrade predictive accuracy and uncertainty calibration. Additionally, the formulation in [Cohen et al. \(2020\)](#) requires all experts to share the same kernel specification and hyperparameters. While this constraint regularises the model and reduces overfitting, it also limits expressiveness, making the model less effective in heteroskedastic and non-stationary settings ([Cao and Fleet 2015](#)).

Related work on information gain in GP models is primarily found in Bayesian optimisation (BO). [Srinivas et al. \(2010, 2012\)](#) used maximum information gain to derive confidence bounds for global GP models in sequential BO. [Desautels et al. \(2014\)](#) employed information gain to address underestimated variances in global GP models used in asynchronous batch BO settings.

Motivated by these developments, we propose a method that leverages information gain bounds (specifically for the Matérn kernel) to address variance overestimation in product-of-experts GP models. Given M local GP experts trained on data subsets, the key challenge is to adjust their posterior variances such that the aggregated predictive variance accurately reflects the global data distribution. To this end, we define a pointwise calibration ratio that can be applied to each local expert at test time using only quantities available in standard GP-pro predictive equations. This formulation establishes the problem setting addressed by our method.

4 Analysing Uncertainty Quantification in Product-of-Experts Gaussian Process Models

In a GP-pro model with M local GPs, each local expert $GP^{(m)}$ is trained on a subset of the data $\mathcal{D}^{(m)}$, while the remaining data $\mathcal{D} \setminus \mathcal{D}^{(m)}$ are effectively unavailable to that expert. Training on smaller subsets reduces computational cost; however, it introduces an important side effect: the predictive variance $\sigma_m^2(\mathbf{x})$ is systematically overestimated (i.e., overly conservative). This follows from the property that, for any fixed input $\mathbf{x}$, the GP posterior variance is non-increasing with respect to the number of training points ([Desautels et al. 2014](#)). In other words, as more training data are observed, the shape of f is better understood, leading to reduced uncertainty.

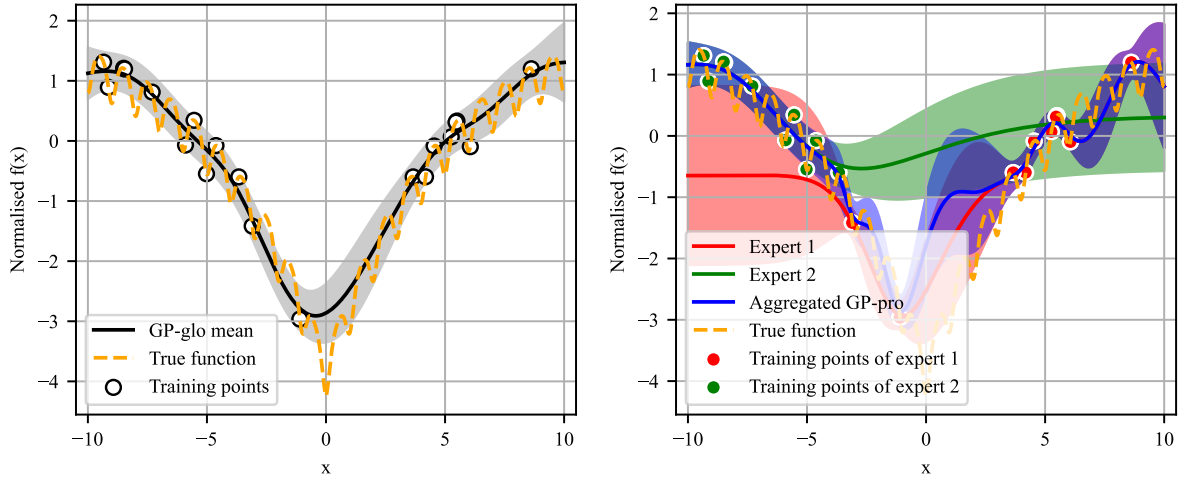


Fig. 1. Illustration of posterior predictions produced by a single global GP model (GP-glo) and a product-of-experts GP model (GP-pro) for the one-dimensional Ackley function. The white dots represent the training data points. The plot on the left shows the posterior prediction of the GP-glo model, with the shaded region indicating the standard deviation at each input value x . The plot on the right shows a GP-pro model comprising two local experts. The balltree assignment method is used to allocate training data points to the local GP models. The green and red curves, together with their corresponding shaded regions, represent the posterior predictions and standard deviations of the two local GP models. The GP-pro model employs an entropy-weighted aggregation scheme to combine the predictions of the local GP models. The blue curve and shaded region represent the aggregated posterior prediction and its associated standard deviation. Compared with the GP-glo model, the GP-pro model exhibits overestimated posterior variance, as evidenced by the wider blue uncertainty intervals.

When aggregating predictions across all local experts $GP^{(m)}$, these inflated variances propagate into the combined posterior (see Figure 1). Consequently, the aggregated predictive variance $\sigma^2(\mathbf{x})$ remains overly conservative, and the posterior reflected by $\sigma^2(\mathbf{x})$ is ‘conservative’ about the algorithm’s actual state of knowledge of the function. Figure 1 illustrates this effect.

To quantitatively evaluate the predictive accuracy and uncertainty quality of the GP-pro model, this study uses three metrics: negative log-likelihood (NLL), root mean squared error (RMSE) and expected normalised calibration error (ENCE). These metrics have been proposed and used by researchers (Levi et al. 2020; Tran et al. 2020) in the literature on uncertainty quantification.

We use root mean squared error (RMSE) to measure point prediction accuracy. RMSE is used because it is sensitive to outliers and is therefore a good measure of the worst-case accuracy (Tran et al. 2020).

To assess uncertainty calibration, we adopt the regression calibration definition of Levi et al. (2020). Let $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ denote the predicted mean and variance. A model is well calibrated if:

$$\forall \sigma : \mathbb{E}_{\mathbf{x}, y} [(\mu(\mathbf{x}) - y)^2 \mid \sigma^2(\mathbf{x}) = \sigma^2] = \sigma^2$$

This condition states that, for predictions with a given uncertainty level, the expected squared error should match the predicted variance σ^2 .

In practice, calibration is evaluated via binning. Let σ_t be the standard deviation of the predicted output probability density function (PDF) p_t , and assume that the examples are ordered by increasing values of σ_t , and the number of bins, J , divides the number of examples, T . The indices of the examples are divided into J bins,

$\{B_j\}_{j=1}^J$, such that $B_j = \{(j-1) \cdot \frac{T}{J} + 1, \dots, j \cdot \frac{T}{J}\}$. Each bin therefore corresponds to an interval in the standard deviation axis: $[\min_{t \in B_j} \{\sigma_t\}, \max_{t \in B_j} \{\sigma_t\}]$. The intervals are non-overlapping, and their boundary values are increasing. Following [Levi et al. \(2020\)](#), we use $J = 10$ bins.

For each bin j , we compute the root of the mean variance (RMV) and the empirical root mean square error (RMSE):

$$\begin{aligned} RMV(j) &= \sqrt{\frac{1}{|B_j|} \sum_{t \in B_j} \sigma_t^2} \\ RMSE(j) &= \sqrt{\frac{1}{|B_j|} \sum_{t \in B_j} (y_t - \hat{y}_t)^2} \end{aligned}$$

where $\hat{y}_t$ is the predicted mean. The expected normalised calibration error (ENCE) ([Levi et al. 2020](#)) is then defined as:

$$ENCE = \frac{1}{J} \sum_{j=1}^J \frac{|RMV(j) - RMSE(j)|}{RMV(j)}$$

ENCE averages the calibration error in each bin, normalised by the bin's mean predicted variance, since a larger variance is expected to have larger errors.

For visual assessment, we use reliability diagrams ([Levi et al. 2020](#)), which plot $RMSE(j)$ against $RMV(j)$, as shown in Figure 2. The diagram is to show that for a perfectly calibrated model, the RMV and the observed RMSE should be approximately equal; hence, the plot should be close to the identity function (i.e., the ideal diagonal line).

We also evaluate models using the negative log-likelihood (NLL) on the test set:

$$NLL = -\frac{1}{N} \sum_{i=1}^N \log p(y_i | \mathcal{N}(\hat{y}_i, \hat{\sigma}_i^2))$$

where y_i is the true value, $\hat{y}_i$ and $\hat{\sigma}_i^2$ are the predicted mean and variance, and N is the number of test points. Lower NLL indicates better overall performance. NLL is used because it provides an overall assessment that captures both predictive accuracy and uncertainty quality ([Tran et al. 2020](#)).

In summary, RMSE evaluates point prediction accuracy, NLL jointly measures accuracy and uncertainty quality, and ENCE specifically assesses calibration.

Table 1 and Figure 2 compare GP-pro, GP-pro-full, and GP-full on four synthetic functions: 1d Ackley, 10d Ackley, 10d Levy, and 10d Rastrigin. The GP-pro model uses balltree assignment, entropy-based weighting, and 200 data points per expert. The GP-pro-full model is a diagnostic baseline in which each expert is trained on the full dataset. GP-full denotes a single global GP. Table 1 lists the performance metrics, Figure 2 shows the reliability diagrams for the GP models, and Table 2 lists the computational overheads incurred by the models.

The results show that all models achieve similar RMSE values. However, GP-pro exhibit higher NLL and ENCE than GP-pro-full and GP-full. As expected, GP-pro-full matches GP-full, since all experts are trained on identical data and therefore produce identical posteriors. As shown in the reliability diagrams in Figure 2, the GP-pro-full and GP-full models exhibit curves that approach the ideal diagonal line, whereas the GP-pro models shows curves that clearly deviate from the diagonal line. These results indicate overestimated posterior variances in the GP-pro model.

Although GP-pro-full provides well-calibrated uncertainty, it incurs significantly higher computational cost. Specifically, GP-pro scales as $\mathcal{O}(M\tilde{n}^3)$, GP-full as $\mathcal{O}(n^3)$, and GP-pro-full as $\mathcal{O}(Mn^3)$, where M is the number of experts, $\tilde{n}$ is the number of data points per expert, n is the size of the training set and $\tilde{n} \ll n$.

Table 1. Average RMSE, NLL, and ENCE results for the GP-pro, GP-pro-full, and GP-full models on four synthetic functions. GP-pro-full and GP-full yield identical results, as expected, since GP-pro-full trains each expert on the entire training dataset, resulting in identical posteriors whose product recovers the global GP solution. GP-pro exhibits higher NLL and ENCE than GP-pro-full and GP-full, indicating variance overestimation in the GP-pro model.

Dataset	Size	Dim	GP-pro-bt-ent			GP-pro-full			GP-full		
			RMSE↓	NLL↓	ENCE↓	RMSE↓	NLL↓	ENCE↓	RMSE↓	NLL↓	ENCE↓
Ackley	500	1	0.300	0.267	0.296	0.290	0.204	0.252	0.290	0.204	0.252
Ackley	5000	10	0.487	0.775	0.245	0.468	0.663	0.057	0.468	0.663	0.057
Levy	5000	10	0.681	1.021	0.160	0.653	0.982	0.137	0.653	0.982	0.137
Rastrigin	5000	10	0.745	1.138	0.127	0.713	1.081	0.039	0.713	1.081	0.039
Average ↓			0.553	0.800	0.207	0.531	0.732	0.121	0.531	0.732	0.121

Table 2. Average training and test times (in seconds) across four synthetic functions for the GP-pro, GP-pro-full, and GP-full models.

Dataset	Size	Dim	GP-pro	GP-pro-full	GP-full
			training, test time	training, test time	training, test time
Ackley	500	1	0.666, 0.018	0.677, 0.021	0.569, 0.013
Ackley	5000	10	3.101, 1.425	963.947, 13.745	47.905, 0.885
Levy	5000	10	3.380, 1.450	982.690, 14.161	46.986, 0.892
Rastrigin	5000	10	3.119, 1.452	972.641, 14.850	50.469, 1.053
Average ↓			2.566, 1.086	729.988, 10.694	36.482, 0.711

Overall, this analysis shows that GP-pro successfully reduces computational complexity but suffers from systematic overestimation of predictive variance, motivating the need for calibration.

5 Information-Based Uncertainty Calibration in Product-of-Experts Gaussian Process Models

To calibrate uncertainty in GP-pro models, the central challenge is to quantify the appropriate magnitude of variance reduction. We propose an information-based variance calibration method that expresses this reduction in terms of conditional mutual information.

Let $n = |\mathcal{D}|$ and $n_m = |\mathcal{D}^{(m)}|$. For a local expert $GP^{(m)}$ and any input $\mathbf{x}$, a natural measure of variance overestimation is the ratio $\frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})}$, where $\sigma_{m,1:n}(\mathbf{x})$ denotes the posterior standard deviation conditioned on the full dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n}$, and $\sigma_{m,1:n_m}(\mathbf{x})$ denotes the posterior standard deviation conditioned on the local subset $\mathcal{D}^{(m)} = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n_m}$. In a local $GP^{(m)}$, the subset of data not assigned to $GP^{(m)}$ is denoted by $\mathcal{D} \setminus \mathcal{D}^{(m)} = \{(\mathbf{x}_i, y_i)\}_{n_m+1 \leq i \leq n}$. The ratio $\frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})}$ is related to $I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m})$, that is, the conditional mutual information with respect to $f(\mathbf{x})$ resulting from observations $\mathbf{y}_{m,n_m+1:n}$, given previous observations $\mathbf{y}_{m,1:n_m}$. The relation is shown below.

LEMMA 5.1. *The ratio of the standard deviation of the posterior over $f(\mathbf{x})$, conditioned on $\mathcal{D}$, to that conditioned on $\mathcal{D}^{(m)}$ is:*

$$\frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})} = \exp \left(-I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) \right) \quad (15)$$

where $\exp$ is the natural exponential function.

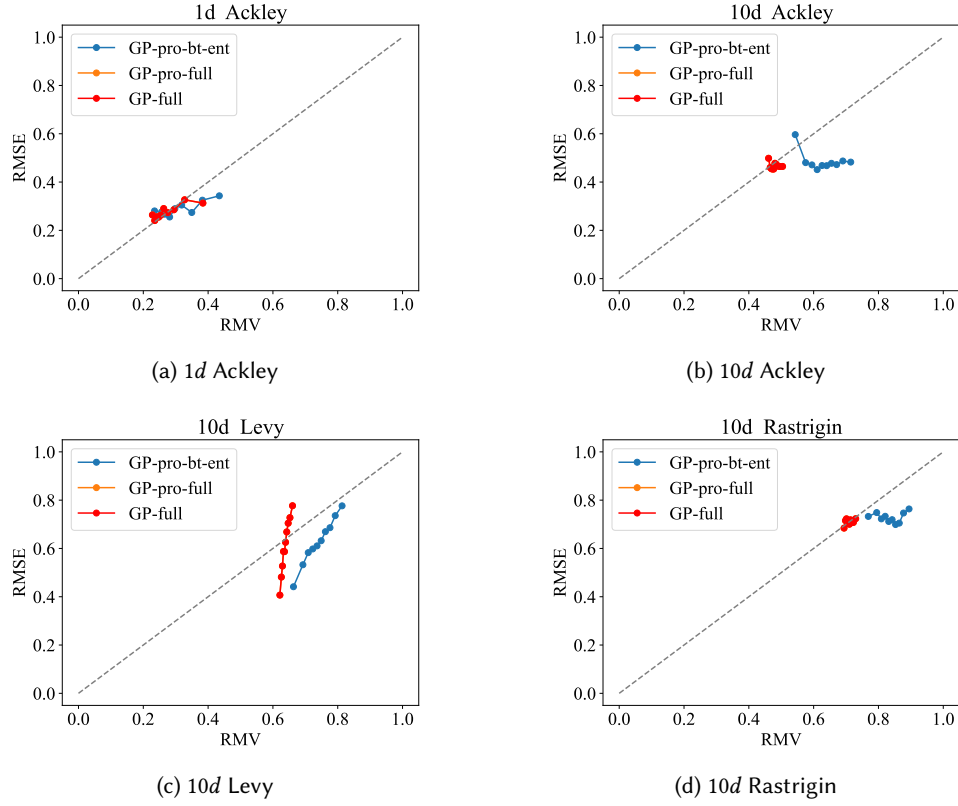


Fig. 2. Reliability diagrams for the GP-pro, GP-pro-full, and GP-full models on four synthetic functions. Well-calibrated models yield curves closer to the ideal diagonal line. The curves of GP-pro-full overlap with those of GP-full, as expected. For the 10d Levy function, GP-full also exhibits noticeable deviation from the diagonal line, which can be attributed to kernel misspecification and the high-dimensional, nonstationary nature of the function. In contrast, GP-pro shows consistently larger deviations across all functions, indicating systematic overestimation of posterior variances in the GP-pro models.

PROOF. Since Gaussian process regression yields Gaussian posterior distributions, the conditional distribution $f(\mathbf{x})|\mathbf{y}$ is Gaussian with variance given by the GP posterior variance. The differential entropy of a univariate Gaussian random variable with variance σ^2 is $\frac{1}{2} \log(2\pi e \sigma^2)$. The lemma therefore follows from:

$$\begin{aligned}
 -I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) &= -(H(f(\mathbf{x}) | \mathbf{y}_{m,1:n_m}) - H(f(\mathbf{x}) | \mathbf{y}_{m,1:n_m}, \mathbf{y}_{m,n_m+1:n})) \\
 &= -(H(f(\mathbf{x}) | \mathbf{y}_{m,1:n_m}) - H(f(\mathbf{x}) | \mathbf{y}_{m,1:n})) \\
 &= -\left(\frac{1}{2} \log(2\pi e \sigma_{m,1:n_m}^2(\mathbf{x})) - \frac{1}{2} \log(2\pi e \sigma_{m,1:n}^2(\mathbf{x})) \right) \\
 &= \log \frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})}
 \end{aligned}$$

□

The ratio $\frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})}$ satisfies:

$$0 < \frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})} = \exp \left(-I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) \right) \leq 1 \quad (16)$$

where the upper bound 1 corresponds to no calibration. As proved by [Dorard \(2012\)](#) and [Srinivas et al. \(2012\)](#) (see Section 2.2), $I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m})$ can be calculated as the sum of the marginal conditional information gain of observations $\mathbf{y}_{m,n_m+1:n}$:

$$I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) = \frac{1}{2} \sum_{i=n_m+1}^n \log \left(1 + \frac{\sigma_{m,i-1}^2(\mathbf{x})}{\ddot{\sigma}_m^2} \right) \quad (17)$$

Evaluating $I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m})$ exactly would require access to the full dataset, which would defeat the purpose of the distributed GP-pro framework whose objective is to avoid the cubic computational complexity of global GP inference. To obtain a tractable approximation, we exploit the monotonicity and submodularity of information gain in GP. In particular, the cumulative information gain from the unseen observations can be lower-bounded by a scaled version of the largest marginal contribution. Let $\Delta_i = \frac{1}{2} \log \left(1 + \frac{\sigma_{m,i-1}^2(\mathbf{x})}{\ddot{\sigma}_m^2} \right)$, then

$$I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) = \sum_{i=n_m+1}^n \Delta_i$$

In GP, information gain is submodular ([Srinivas et al. 2012](#)). Submodularity implies diminishing returns, i.e. the marginal gains are nonincreasing as additional observations are incorporated:

$$\Delta_{n_m+1} \geq \Delta_{n_m+2} \geq \dots \geq \Delta_n$$

Thus the largest marginal gain occurs at the first element, i.e., $\max_{i \in n_m+1:n} \Delta_i = \Delta_{n_m+1}$.

Since all marginal gains are nonnegative, this yields the conservative lower bound:

$$I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) \geq \frac{e-1}{e} \frac{1}{2} \log \left(1 + \frac{\sigma_{m,n_m}^2(\mathbf{x})}{\ddot{\sigma}_m^2} \right) \quad (18)$$

Although Equation (13) is commonly derived in the context of greedy subset selection, we use it here only as a general inequality for monotone submodular functions to obtain a conservative lower bound on cumulative information gain. The derivation therefore does not assume that the dataset itself is generated by a greedy algorithm.

Substituting this bound into the variance ratio expression in Lemma 5.1 yields:

$$\frac{\sigma_{m,1:n}(\mathbf{x})}{\sigma_{m,1:n_m}(\mathbf{x})} = \exp \left(-I(f(\mathbf{x}); \mathbf{y}_{m,n_m+1:n} | \mathbf{y}_{m,1:n_m}) \right) \leq \exp \left(-\frac{e-1}{e} \frac{1}{2} \log \left(1 + \frac{\sigma_{m,n_m}^2(\mathbf{x})}{\ddot{\sigma}_m^2} \right) \right) \quad (19)$$

Note that the posterior variance $\sigma_m^2(\mathbf{x})$ computed by the local expert $GP^{(m)}$ is exactly the variance conditioned on its local dataset $\mathcal{D}^{(m)}$, i.e., $\sigma_m^2(\mathbf{x}) = \sigma_{m,n_m}^2(\mathbf{x})$. We therefore use the shorter notation $\sigma_m^2(\mathbf{x})$ in the following derivation.

We define the calibration ratio for any $\mathbf{x} \in \mathcal{X}$ at a local $GP^{(m)}$ as:

$$\alpha_m(\mathbf{x}) = \exp \left(-\frac{e-1}{e} \frac{1}{2} \log \left(1 + \frac{\sigma_m^2(\mathbf{x})}{\ddot{\sigma}_m^2} \right) \right) \quad (20)$$

The proposed calibration ratio does not reduce the posterior variance more aggressively than the reduction implied by the lower bound on conditional mutual information. As a result, the calibrated variance

$$\hat{\sigma}_m^2(\mathbf{x}) = \alpha_m(\mathbf{x}) \cdot \sigma_m^2(\mathbf{x}) \quad (21)$$

remains a conservative estimate of the variance that would be obtained if the additional observations in $\mathcal{D} \setminus \mathcal{D}^{(m)}$ were incorporated.

The calibration ratio $\alpha_m(\mathbf{x})$ implicitly captures the potential contribution of the missing global data while remaining computable using only quantities available to the local expert, i.e., the posterior variance $\sigma_m^2(\mathbf{x})$, which is already computed as part of the standard GP-pro predictive equations. Consequently, the calibration ratio $\alpha_m(\mathbf{x})$ can be evaluated efficiently at test time for arbitrary inputs $\mathbf{x}$ without requiring access to other experts' datasets or additional training-time computation. In this sense, the posterior variance predicted by the local expert serves as a compact summary of the information already incorporated by $GP^{(m)}$, while the bound provides a conservative estimate of the additional information that could be contributed by the missing observations.

The factor $\frac{e-1}{e}$ introduces a controlled degree of conservativeness. In the context of uncertainty calibration, this conservativeness is intentional: an overly aggressive reduction of posterior variance can lead to numerically unstable predictions. The proposed bound therefore acts as a safeguard against over-calibration. While tighter, data-dependent bounds on conditional information gain may further improve calibration accuracy, exploring such alternatives would require additional assumptions or adaptive estimation strategies and is left for future work. Empirically, the current bound already yields consistent improvements in NLL and ENCE across diverse datasets, as reported in Section 6.

In the proposed calibration rule, the information gain term corresponds to the largest marginal information gain contributed by a single observation, as discussed above. In practice, we normalise this quantity so that its magnitude does not exceed one. This choice is consistent with the analysis of [Srinivas et al. \(2012\)](#), which shows that under standard assumptions with unit prior variance, the marginal information gain contributed by a single observation is bounded by a constant of order one for commonly used kernels such as the Matérn kernel. Since our calibration uses only the largest marginal information gain term, adopting a normalised scale with an upper bound of one provides a convenient and stable numerical reference without affecting the functional form of the calibration rule. Although the derivation relies on standard assumptions regarding normalised information gain and the reference to Matérn kernels, the calibration itself depends only on posterior variances and noise levels that are already available at test time, and does not otherwise rely on kernel-specific properties. We therefore expect the method to extend to other commonly used stationary kernels, although a formal analysis of such extensions is left for future work.

Aggregating posterior predictions. During test time, at a candidate input $\mathbf{x}'$, each local $GP^{(m)}$ independently computes the posterior predictive mean $\mu_m(\mathbf{x}')$ and variance $\sigma_m^2(\mathbf{x}')$ using the standard GP-pro predictive equations. For completeness, we restate the standard GP-pro predictive equations below:

$$\begin{aligned} \mu_m(\mathbf{x}') &= \tilde{\mu}^{(m)}(\mathbf{x}') + \mathbf{k}^{(m)}(\mathbf{x}')^\top \left(\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I} \right)^{-1} \left(\mathbf{y}^{(m)} - \tilde{\mu}^{(m)} \right) \\ \sigma_m^2(\mathbf{x}') &= k^{(m)}(\mathbf{x}', \mathbf{x}') - \mathbf{k}^{(m)}(\mathbf{x}')^\top \left(\mathbf{K}^{(m)} + \ddot{\sigma}_m^2 \mathbf{I} \right)^{-1} \mathbf{k}^{(m)}(\mathbf{x}') \end{aligned}$$

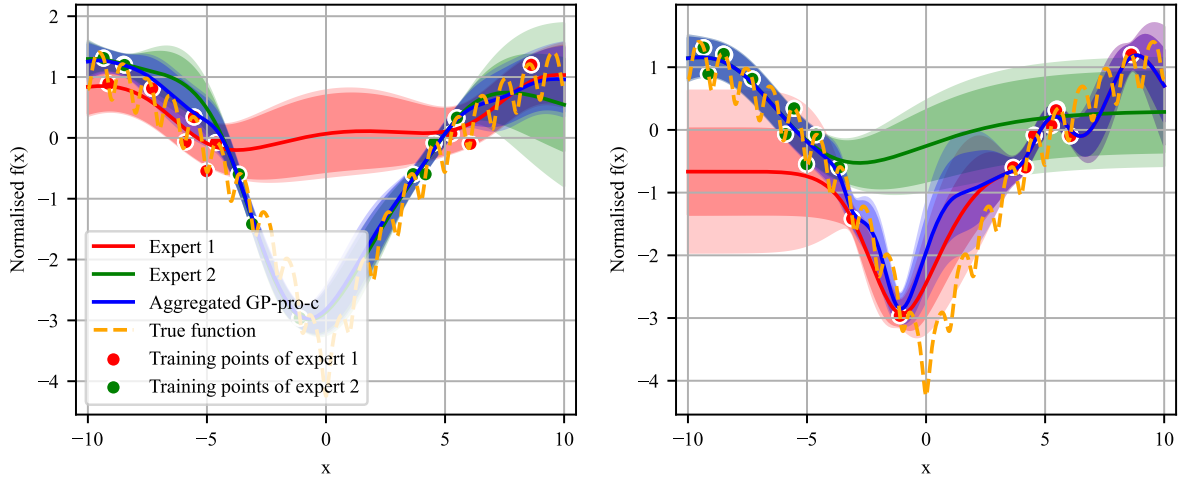


Fig. 3. Illustration of posterior predictions before and after applying the calibration ratio to GP-pro-c models comprising two local experts for the one-dimensional Ackley function. The GP-pro-c model on the left uses the random assignment method, whereas the model on the right uses balltree assignment. Both GP-pro-c models employ an entropy-weighted aggregation scheme. The posterior predictions of the two local GP models are represented by the green and red curves, with the corresponding light green and light red shaded regions indicating the standard deviation at each input value x . The blue curve and shaded region represent the aggregated posterior prediction. The darker green, red, and blue shaded regions indicate the reduced standard deviations obtained after applying the calibration ratio.

With the calibrated variance $\hat{\sigma}_m^2(\mathbf{x}') = \alpha_m(\mathbf{x}') \cdot \sigma_m^2(\mathbf{x}')$, the aggregated predictive mean μ and variance σ^2 at the candidate $\mathbf{x}'$ are derived as:

$$\mu(\mathbf{x}') = \sigma^2(\mathbf{x}') \sum_{m=1}^M \left(w_m(\mathbf{x}') \mu_m(\mathbf{x}') \frac{1}{\hat{\sigma}_m^2(\mathbf{x}')} \right) \quad (22)$$

$$\sigma^2(\mathbf{x}') = \left(\sum_{m=1}^M \left(w_m(\mathbf{x}') \frac{1}{\hat{\sigma}_m^2(\mathbf{x}')} \right) \right)^{-1} \quad (23)$$

Although Equations (22) and (23) retain the algebraic structure of the standard GP-pro aggregation, the predictive uncertainty entering the aggregation is fundamentally different. In the proposed model, each expert contributes a calibrated posterior variance $\hat{\sigma}_m^2(\mathbf{x}')$ obtained from the information-based shrinkage rule, rather than the raw posterior variance produced by the local GP. This modification corrects the systematic variance inflation that arises when experts are trained on disjoint subsets of data. Importantly, the aggregation rule itself remains unchanged, which preserves the computational efficiency and scalability of the original GP-pro framework while improving the reliability of its uncertainty estimates.

Figure 3 illustrates the effect of calibration on posterior predictions for the Ackley function using GP-pro-c models comprising two local experts. The calibrated model reduces predictive variance while maintaining accurate mean predictions.

Table 3 summarises training and test-time complexity for global GP, vanilla GP-pro and GP-pro-c with variance calibration. GP-pro-c retains efficient product-of-experts aggregation with calibrated posterior variances. Each local GP requires $O(n_m^3)$ training, and test-time mean and variance computations are $O(n_m)$ and $O(n_m^2)$ per

Table 3. Computational complexity comparison.

Method	Training complexity	Test-time mean	Test-time variance	Notes
Global GP	$O(n^3)$	$O(n)$	$O(n^2)$	Full dataset, accurate variance
Vanilla GP-pro (M experts, n_m points each)	$O(Mn_m^3)$	$O(Mn_m)$	$O(Mn_m^2)$	Distributed experts, standard PoE
GP-pro-c (proposed)	$O(Mn_m^2)$	$O(Mn_m)$	$O(Mn_m^2)$	Adds calibrated variance via $\alpha_m(\mathbf{x})$ with negligible cost

expert, respectively. The calibration ratio $\alpha_m(\mathbf{x})$ is computed directly from the already-available posterior variance, introducing negligible extra computation, and preserving the efficiency advantage over a global GP trained on all n points.

6 Experiments

This section presents the empirical evaluation of the GP-pro-c model. The implementation is publicly available at <https://github.com/yhong123/gp-pro-c>.

All experiments were conducted using GPyTorch (Gardner et al. 2018) for GP regression. Prior to model fitting, input features were rescaled to $[0, 1]^d$, and outputs were standardised. Each GP was parameterised with a Matérn-5/2 covariance function with automatic relevance determination (ARD) and a constant mean function. Hyperparameters were learned by maximising the log marginal likelihood.

Experiments were conducted on four synthetic functions with known global minima: the 10d Ackley (5000 data points), 10d Levy (5000 data points), 10d Rastrigin (5000 data points) and 50d Rosenbrock (50000 data points) functions. All function values were corrupted with additive Gaussian noise $\mathcal{N}(0, 0.25)$.

We also consider six real-world regression datasets: Concrete (8d feature space, 1030 data points), Airfoil (5d, 1503 data points), Power (4d, 9568 data points), Parkinson (20d, 5875 data points), Kinetic (8d, 40000 data points) and Protein (9d, 45730 data points).

Each experiment on a function or dataset was repeated ten times with different random seeds. In each run, the data points in a function or dataset were partitioned such that 50% were used as training points and 50% as test points.

The performance of each GP model was measured using three performance metrics: root mean squared error (RMSE), expected normalised calibration error (ENCE) and negative log-likelihood (NLL). Details of these metrics are provided in Section 4.

All experiments were conducted on a single Apple M2 processor (3.49 GHz) with 8 GB of memory.

6.1 Comparing Data Assignment Methods and Weighting Schemes

The first set of experiments examines the impact of the proposed information-based calibration across variants of GP-pro-c models using different data assignment methods and weighting schemes. Specifically, we consider random, k -means, and balltree assignment methods, combined with entropy-, variance-, and uniform-weighted schemes, with 200 data points per expert. These models are compared against corresponding GP-pro variants without calibration.

Table 4 summarises the model configurations and their labels. Each model name (e.g., rd-ent-c) encodes its design choices: ‘rd’, ‘kx’, and ‘bt’ denote random, k -means, and balltree assignment methods, respectively, while ‘ent’, ‘var’, and ‘uni’ denote entropy-, variance-, and uniform-weighted schemes. The suffix ‘c’ indicates the use of information-based variance calibration.

Tables 5-7 report the performance metrics for each GP model. Overall, across all data assignment methods and weighting schemes, GP-pro-c achieves RMSE values comparable to those of GP-pro. However, GP-pro-c

Table 4. Summary of the models used in the experiments for the comparison and evaluation of the GP-pro-c and GP-pro models.

Model	Data assignment method	Weighting scheme	Uncertainty calibration
bt-ent	balltree	entropy	no
bt-var	balltree	variance	no
bt-uni	balltree	uniform	no
kx-ent	<i>k</i> -means	entropy	no
kx-var	<i>k</i> -means	variance	no
kx-uni	<i>k</i> -means	uniform	no
rd-ent	random	entropy	no
rd-var	random	variance	no
rd-uni	random	uniform	no
bt-ent-c	balltree	entropy	yes
bt-var-c	balltree	variance	yes
bt-uni-c	balltree	uniform	yes
kx-ent-c	<i>k</i> -means	entropy	yes
kx-var-c	<i>k</i> -means	variance	yes
kx-uni-c	<i>k</i> -means	uniform	yes
rd-ent-c	random	entropy	yes
rd-var-c	random	variance	yes
rd-uni-c	random	uniform	yes

consistently attains lower NLL and ENCE scores, indicating improved uncertainty calibration. These results suggest that the proposed information-based calibration effectively mitigates overestimation of predictive variance in local experts, regardless of the assignment or weighting method.

Among GP-pro-c variants, the combination of balltree assignment and entropy-based weighting performs best, achieving the lowest RMSE, NLL, and ENCE. This observation is consistent with prior findings by [Cao and Fleet \(2014, 2015\)](#) for GP-pro models without calibration.

The balltree and *k*-means assignment methods partition training data points based on similarity measures. Data points with close similarities are grouped and used to fit individual local GP models. These lead to better modelling of the input domain and yield better performance results in comparison to the random assignment method. Between the balltree and *k*-means, the balltree method is more beneficial because the allocation is done so that each local GP model has a nearly equivalent number of training data points. In contrast, by using the *k*-means assignment, each local GP model can be allocated a different number of data points. This may result in an undesirable condition where some local GPs have few data points while others are allocated a large number of data points, which will incur higher computational overheads.

The choice of assignment method influences predictive accuracy and uncertainty calibration in distinct ways. Random assignment constructs weakly correlated local GP experts, leading to conservative predictive variances that tend to match empirical errors and therefore achieve lower ENCE. However, the lack of spatial locality degrades posterior mean estimates, resulting in higher RMSE and NLL. In contrast, balltree and *k*-means assignments preserve locality and improve predictive accuracy, but induce greater information redundancy across experts. This redundancy amplifies variance misestimation in the uncalibrated GP-pro model, making uncertainty calibration more challenging and yielding higher ENCE despite superior predictive performance. These observations highlight that ENCE should be interpreted jointly with RMSE and NLL rather than in isolation.

It is worth noting that lower absolute ENCE values do not necessarily imply larger uncertainty corrections. While random assignment achieves consistently lower ENCE due to weaker correlations among local experts, locality-aware assignments such as balltree and *k*-means introduce greater information redundancy, resulting in

Table 5. Average RMSE values across four synthetic functions and six benchmark datasets using balltree, k -means, and random assignment methods, as well as entropy-, variance-, and uniform-weighted schemes for the GP-pro-c and GP-pro models. GP-pro-c models with information-based calibration achieve RMSE values comparable to those of GP-pro models, indicating that information-based calibration improves uncertainty quantification without affecting predictive accuracy.

Dataset	Size	Dim	bt-ent-c	bt-var-c	bt-uni-c	kx-ent-c	kx-var-c	kx-uni-c	rd-ent-c	rd-var-c	rd-uni-c
Ackley	5000	10	0.487	0.490	0.490	0.497	0.502	0.500	0.471	0.471	0.471
Levy	5000	10	0.681	0.685	0.697	0.679	0.681	0.686	0.670	0.670	0.670
Rastrigin	5000	10	0.745	0.745	0.755	0.746	0.747	0.751	0.737	0.736	0.738
Rosenbrock	50000	50	0.610	0.612	0.624	0.691	0.687	0.690	0.586	0.584	0.586
Concrete	1030	8	0.336	0.340	0.338	0.331	0.335	0.338	0.352	0.353	0.352
Airfoil	1503	5	0.283	0.291	0.295	0.294	0.296	0.310	0.314	0.316	0.317
Parkinsons	5875	20	0.027	0.031	0.094	0.042	0.063	0.172	0.042	0.047	0.048
Power	9568	4	0.249	0.263	0.312	0.252	0.269	0.488	0.245	0.245	0.245
Kin40K	40000	8	0.384	0.457	0.500	0.365	0.452	0.503	0.466	0.474	0.486
Protein	45730	9	0.664	0.688	0.888	0.725	0.704	1.021	0.813	0.815	0.814
Average RMSE ↓			0.447	0.460	0.499	0.462	0.474	0.546	0.470	0.471	0.473

(a) GP-pro-c models

Dataset	Size	Dim	bt-ent	bt-var	bt-uni	kx-ent	kx-var	kx-uni	rd-ent	rd-var	rd-uni
Ackley	5000	10	0.487	0.490	0.489	0.497	0.502	0.500	0.471	0.471	0.471
Levy	5000	10	0.681	0.684	0.695	0.679	0.680	0.685	0.670	0.670	0.670
Rastrigin	5000	10	0.745	0.745	0.753	0.746	0.747	0.750	0.737	0.736	0.738
Rosenbrock	50000	50	0.610	0.612	0.623	0.693	0.689	0.692	0.587	0.585	0.587
Concrete	1030	8	0.336	0.340	0.338	0.331	0.335	0.337	0.352	0.353	0.352
Airfoil	1503	5	0.286	0.294	0.297	0.293	0.295	0.307	0.314	0.316	0.317
Parkinsons	5875	20	0.027	0.031	0.073	0.040	0.059	0.140	0.042	0.047	0.048
Power	9568	4	0.249	0.262	0.309	0.252	0.268	0.484	0.245	0.245	0.245
Kin40K	40000	8	0.375	0.446	0.486	0.352	0.436	0.484	0.464	0.472	0.484
Protein	45730	9	0.663	0.687	0.886	0.727	0.703	1.027	0.813	0.815	0.814
Average RMSE ↓			0.446	0.459	0.495	0.461	0.471	0.541	0.470	0.471	0.473

(b) GP-pro models

more severe variance overestimation in the uncalibrated GP-pro model. As a result, the proposed information-based calibration produces larger uncertainty reductions for balltree and k -means assignments, even though their final ENCE values may remain higher than those of random assignment.

The entropy-weighted scheme computes the weighting factors of each local GP model based on the entropy difference between the prior and posterior distributions at a test point. [Cao and Fleet \(2015\)](#) explained that the entropy-weighted scheme is beneficial because maximising the entropy differences corresponds to minimising the KL divergence between the distribution of individual local GPs and the unknown ground-truth aggregating distribution. [Cao and Fleet \(2015\)](#) formulated the product of a collection of local GP models as a logarithmic opinion pool of experts ([Heskes 1997](#)), where the decision of weighting factors for each local GP can be optimised by minimising the KL divergence of the opinion pool, which can be decomposed into the KL divergences of individual local GPs. Within the entropy-weighted scheme, the prior entropy reflects the uncertainty before observing the training data, while the posterior entropy reflects the uncertainty after observing the training data. A large entropy difference corresponds to a small KL divergence and entails a higher reliability of a local GP model. Thus, by using the entropy-weighted scheme, higher weights are given to local GPs that are more reliable (larger entropy difference) in their predictions, and lower weights are given to local GPs that are less reliable (lower entropy difference). These lead to better aggregation of the posterior predictions.

The variance-weighted scheme, which computes the weighting factors based on the variance difference, works in a similar way to the entropy-weighted scheme. A consistent pattern emerges when comparing variance- and

Table 6. Average NLL values across four synthetic functions and six benchmark datasets using balltree, k -means, and random assignment methods, as well as entropy-, variance-, and uniform-weighted schemes for the GP-pro-c and GP-pro models. GP-pro-c models with information-based calibration achieve lower NLL values than GP-pro models, demonstrating that information-based calibration improves uncertainty quantification without affecting predictive accuracy.

Dataset	Size	Dim	bt-ent-c	bt-var-c	bt-uni-c	kx-ent-c	kx-var-c	kx-uni-c	rd-ent-c	rd-var-c	rd-uni-c
Ackley	5000	10	0.742	0.755	0.770	0.769	0.786	0.800	0.691	0.691	0.692
Levy	5000	10	1.012	1.021	1.043	1.015	1.021	1.034	0.983	0.982	0.982
Rastrigin	5000	10	1.124	1.125	1.141	1.127	1.128	1.137	1.118	1.116	1.118
Rosenbrock	50000	50	0.918	0.923	0.944	1.038	1.031	1.035	0.874	0.871	0.874
Concrete	1030	8	0.287	0.304	0.316	0.253	0.269	0.304	0.331	0.337	0.334
Airfoil	1503	5	0.093	0.133	0.189	0.045	0.072	0.137	0.217	0.228	0.234
Parkinsons	5875	20	-1.374	-1.315	-0.730	-1.316	-1.117	-0.331	-1.114	-1.063	-1.057
Power	9568	4	0.034	0.082	0.232	0.038	0.103	0.784	0.015	0.015	0.015
Kin40K	40000	8	0.466	0.626	0.724	0.424	0.622	0.737	0.647	0.662	0.683
Protein	45730	9	0.932	0.989	1.306	1.113	1.018	1.491	1.220	1.222	1.221
Average NLL ↓			0.423	0.464	0.593	0.451	0.493	0.713	0.498	0.506	0.510

(a) GP-pro-c models

Dataset	Size	Dim	bt-ent	bt-var	bt-uni	kx-ent	kx-var	kx-uni	rd-ent	rd-var	rd-uni
Ackley	5000	10	0.775	0.792	0.812	0.804	0.825	0.845	0.707	0.709	0.710
Levy	5000	10	1.021	1.031	1.054	1.031	1.041	1.056	0.980	0.979	0.980
Rastrigin	5000	10	1.138	1.141	1.159	1.147	1.151	1.164	1.113	1.113	1.116
Rosenbrock	50000	50	0.929	0.934	0.953	1.028	1.023	1.027	0.881	0.879	0.882
Concrete	1030	8	0.287	0.305	0.317	0.254	0.271	0.307	0.331	0.337	0.333
Airfoil	1503	5	0.099	0.139	0.196	0.047	0.074	0.141	0.219	0.230	0.236
Parkinsons	5875	20	-1.352	-1.293	-0.748	-1.304	-1.113	-0.410	-1.113	-1.062	-1.056
Power	9568	4	0.034	0.082	0.222	0.038	0.103	0.747	0.015	0.015	0.015
Kin40K	40000	8	0.471	0.631	0.730	0.431	0.629	0.746	0.661	0.676	0.697
Protein	45730	9	0.930	0.988	1.305	1.106	1.013	1.487	1.219	1.221	1.220
Average NLL ↓			0.433	0.475	0.600	0.458	0.502	0.711	0.501	0.510	0.513

(b) GP-pro models

entropy-weighted schemes. Entropy weighting achieves slightly better predictive accuracy, as measured by RMSE and NLL, likely due to its smoother aggregation of local experts. In contrast, variance-based weighting yields marginally lower ENCE, indicating improved uncertainty calibration. This behaviour aligns with the theoretical distinction between the two schemes: variance weighting more strongly penalises uncertain experts, which benefits calibration, while entropy weighting provides a more conservative aggregation that favours predictive accuracy.

In contrast, the uniform-weighted scheme assigns equal weights to all experts and consistently performs worse. This suggests that assuming equal reliability across experts is suboptimal for aggregating posterior predictions.

Figure 4 presents the reliability diagrams for GP-pro-c and GP-pro models under different data assignment methods, including balltree, k -means, and random assignment. Across all cases, GP-pro-c consistently exhibits reliability curves closer to the ideal diagonal line than GP-pro, demonstrating improved calibration. Deviations from the diagonal vary across datasets: smoother or lower-dimensional datasets (e.g., Concrete, Airfoil, Power) show near-diagonal curves, whereas noisier or higher-dimensional datasets (e.g., Levy, Rosenbrock, Parkinsons, Kin40K) exhibit larger residual deviations.

Figure 7 in Appendix B further reports reliability diagrams for GP-pro-c and GP-pro using the balltree assignment while comparing entropy-, variance-, and uniform-weighted aggregation schemes. The results show that, irrespective of the weighting scheme, GP-pro-c achieves curves closer to the diagonal. Among the GP-pro-c

Table 7. Average ENCE values across four synthetic functions and six benchmark datasets using balltree, k -means, and random assignment methods, as well as entropy-, variance-, and uniform-weighted schemes for the GP-pro-c and GP-pro models. GP-pro-c models with information-based calibration achieve lower ENCE values than GP-pro models, demonstrating improved uncertainty quantification.

Dataset	Size	Dim	bt-ent-c	bt-var-c	bt-uni-c	kx-ent-c	kx-var-c	kx-uni-c	rd-ent-c	rd-var-c	rd-uni-c
Ackley	5000	10	0.190	0.207	0.236	0.206	0.224	0.259	0.120	0.122	0.126
Levy	5000	10	0.103	0.114	0.113	0.159	0.174	0.186	0.098	0.099	0.100
Rastrigin	5000	10	0.060	0.064	0.084	0.065	0.070	0.089	0.083	0.079	0.076
Rosenbrock	50000	50	0.116	0.104	0.081	0.188	0.181	0.174	0.203	0.206	0.205
Concrete	1030	8	0.162	0.151	0.178	0.157	0.174	0.217	0.167	0.165	0.162
Airfoil	1503	5	0.190	0.187	0.153	0.140	0.178	0.252	0.151	0.157	0.162
Parkinsons	5875	20	0.750	0.735	0.543	0.635	0.528	0.233	0.689	0.678	0.678
Power	9568	4	0.069	0.081	0.092	0.052	0.103	0.312	0.059	0.059	0.057
Kin40K	40000	8	0.306	0.277	0.268	0.372	0.331	0.311	0.178	0.175	0.169
Protein	45730	9	0.097	0.076	0.072	0.299	0.095	0.210	0.085	0.082	0.081
Average ENCE ↓			0.204	0.200	0.182	0.227	0.206	0.224	0.183	0.182	0.182

(a) GP-pro-c models

Dataset	Size	Dim	bt-ent	bt-var	bt-uni	kx-ent	kx-var	kx-uni	rd-ent	rd-var	rd-uni
Ackley	5000	10	0.245	0.263	0.292	0.260	0.279	0.314	0.166	0.172	0.175
Levy	5000	10	0.160	0.172	0.179	0.214	0.232	0.252	0.103	0.104	0.107
Rastrigin	5000	10	0.127	0.138	0.153	0.153	0.164	0.182	0.064	0.065	0.070
Rosenbrock	50000	50	0.167	0.163	0.145	0.191	0.187	0.180	0.240	0.242	0.240
Concrete	1030	8	0.166	0.150	0.176	0.155	0.178	0.230	0.166	0.164	0.161
Airfoil	1503	5	0.191	0.188	0.155	0.143	0.183	0.263	0.159	0.163	0.166
Parkinsons	5875	20	0.759	0.747	0.657	0.662	0.570	0.400	0.690	0.678	0.678
Power	9568	4	0.071	0.085	0.098	0.059	0.122	0.286	0.059	0.059	0.057
Kin40K	40000	8	0.348	0.326	0.323	0.420	0.391	0.377	0.218	0.217	0.211
Protein	45730	9	0.086	0.074	0.072	0.281	0.086	0.183	0.056	0.053	0.057
Average ENCE ↓			0.232	0.231	0.225	0.254	0.239	0.267	0.192	0.192	0.192

(b) GP-pro models

variants, the combination of balltree assignment and entropy weighting consistently produces reliability curves closest to the diagonal, indicating more accurate uncertainty quantification.

Table 8 lists the training and test time required by the GP-pro-c and GP-pro models. GP-pro-c exhibits computational costs comparable to GP-pro, indicating that the proposed calibration introduces only negligible overhead.

Overall, the results demonstrate that the information-based calibration method improves uncertainty quantification in GP-pro-c without compromising predictive accuracy or computational efficiency.

6.2 Comparing the Number of Training Data Points per Local Gaussian Process Model

The second set of experiments examines the impact of varying the number of training data points per local GP on predictive performance and computational cost of the GP-pro-c models. Specifically, we consider models with 100, 200, 300, 400, 500, and 600 data points per local expert.

As discussed in Section 1, product-of-experts GP models reduce the computational overhead of a single global GP from cubic complexity $O(n^3)$ to $O(M\tilde{n}^3)$, where n is the number of training data points, $\tilde{n}$ is the number of training data points under each local GP, and $\tilde{n} \ll n$. Within a local GP model, the more data points allocated, the more accurate the model is, while also incurring higher computational overhead.

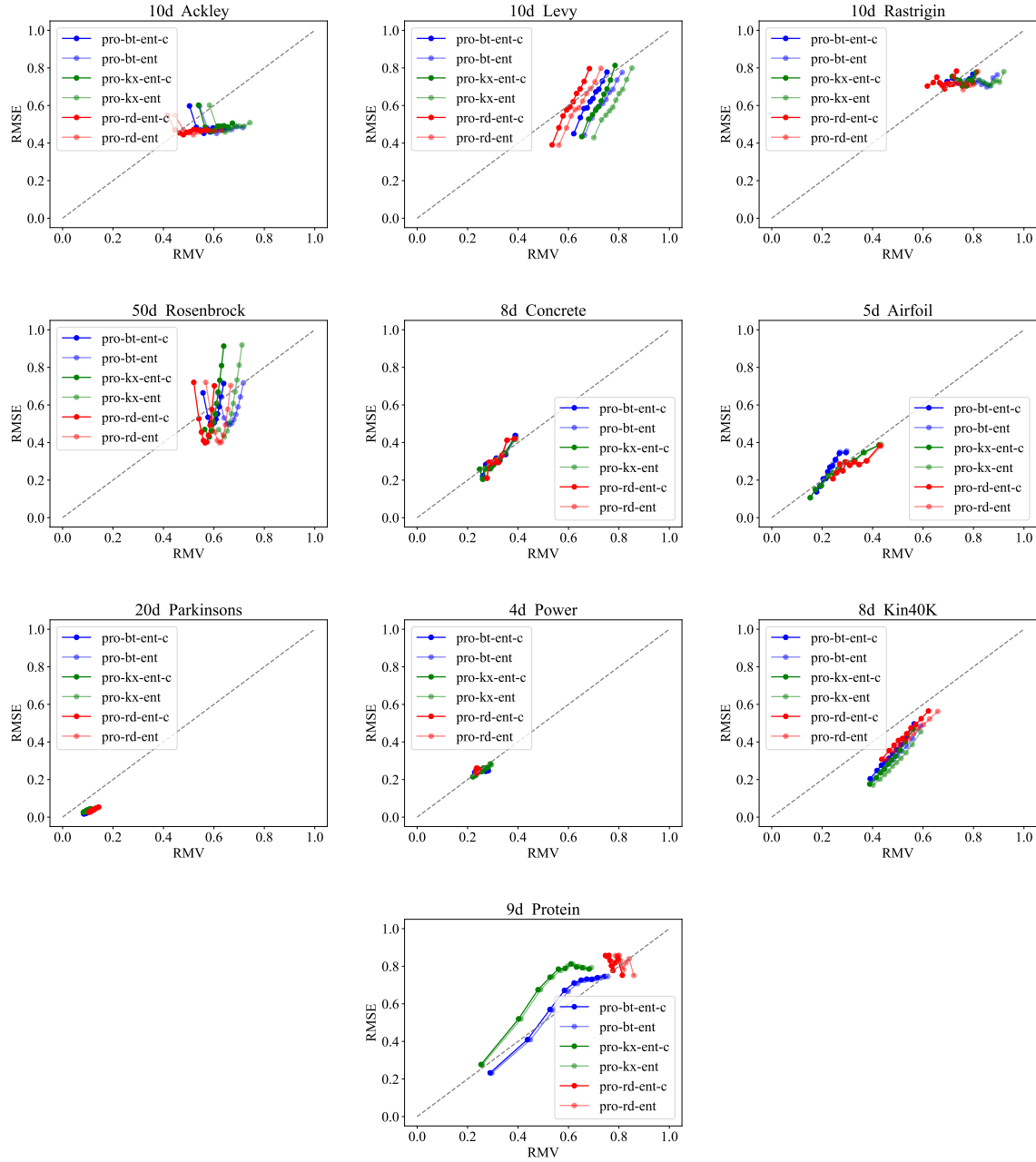


Fig. 4. Reliability diagrams comparing GP-pro-c and GP-pro models using balltree, k -means, and random assignment methods across four synthetic functions and six regression datasets. GP-pro-c consistently yields curves closer to the ideal diagonal than GP-pro, indicating improved uncertainty calibration. The degree of deviation from the diagonal varies across datasets: smoother or lower-dimensional problems (e.g., Concrete, Airfoil, and Power) show near-diagonal curves, whereas noisier or higher-dimensional problems (e.g., Levy, Rosenbrock, Parkinsons, and Kin40K) show larger residual deviations.

Table 8. Average training and test times (in seconds) across four synthetic functions and six regression datasets using balltree, k -means, and random assignment methods, as well as entropy- and variance-weighted schemes (the uniform-weighted scheme is omitted for brevity) for the GP-pro-c and GP-pro models. Variants of GP-pro-c have training and test times similar to those of GP-pro.

Dataset	Size	Dim	bt-ent-c training, test time	bt-var-c	kx-ent-c	kx-var-c	rd-ent-c	rd-var-c
Ackley	5000	10	3.101, 1.427	3.101, 1.427	3.265, 1.396	3.265, 1.395	3.166, 1.427	3.166, 1.427
Levy	5000	10	3.380, 1.452	3.380, 1.452	3.479, 1.352	3.479, 1.352	3.413, 1.459	3.413, 1.459
Rastrigin	5000	10	3.119, 1.454	3.119, 1.454	3.371, 1.435	3.371, 1.435	3.069, 1.449	3.069, 1.449
Rosenbrock	50000	50	32.522, 85.483	32.522, 85.467	33.917, 84.008	33.917, 83.992	30.327, 81.578	30.327, 81.565
Concrete	1030	8	1.263, 0.037	1.263, 0.037	1.238, 0.031	1.238, 0.031	1.301, 0.033	1.301, 0.033
Airfoil	1503	5	1.359, 0.066	1.359, 0.066	1.367, 0.061	1.367, 0.061	1.282, 0.063	1.282, 0.063
Parkinsons	5875	20	4.282, 2.581	4.282, 2.581	4.514, 2.724	4.514, 2.723	3.745, 2.463	3.745, 2.463
Power	9568	4	5.779, 0.538	5.779, 0.537	6.151, 0.578	6.151, 0.577	5.495, 0.502	5.495, 0.502
Kin40K	40000	8	22.336, 41.573	22.336, 41.558	24.403, 42.255	24.403, 42.242	20.518, 41.063	20.518, 41.055
Protein	45730	9	27.192, 58.806	27.192, 58.793	32.110, 63.094	32.110, 63.068	24.176, 58.166	24.176, 58.150
Average time ↓			10.433, 19.342	10.433, 19.337	11.381, 19.693	11.381, 19.688	9.649, 18.820	9.649, 18.816

(a) Training and test time for GP-pro-c models

Dataset	Size	Dim	bt-ent training, test time	bt-var	kx-ent	kx-var	rd-ent	rd-var
Ackley	5000	10	3.101, 1.425	3.101, 1.425	3.265, 1.394	3.265, 1.394	3.166, 1.425	3.166, 1.425
Levy	5000	10	3.380, 1.450	3.380, 1.450	3.479, 1.350	3.479, 1.350	3.413, 1.457	3.413, 1.457
Rastrigin	5000	10	3.119, 1.452	3.119, 1.452	3.371, 1.433	3.371, 1.433	3.069, 1.447	3.069, 1.447
Rosenbrock	50000	50	32.522, 85.375	32.522, 85.356	33.917, 83.896	33.917, 83.879	30.327, 81.470	30.327, 81.457
Concrete	1030	8	1.263, 0.036	1.263, 0.036	1.238, 0.031	1.238, 0.031	1.301, 0.033	1.301, 0.033
Airfoil	1503	5	1.359, 0.065	1.359, 0.065	1.367, 0.060	1.367, 0.060	1.282, 0.062	1.282, 0.062
Parkinsons	5875	20	4.282, 2.579	4.282, 2.579	4.514, 2.721	4.514, 2.721	3.745, 2.461	3.745, 2.460
Power	9568	4	5.779, 0.533	5.779, 0.532	6.151, 0.573	6.151, 0.573	5.495, 0.498	5.495, 0.497
Kin40K	40000	8	22.336, 41.503	22.336, 41.489	24.403, 42.187	24.403, 42.172	20.518, 40.994	20.518, 40.987
Protein	45730	9	27.192, 58.716	27.192, 58.701	32.110, 62.996	32.110, 62.982	24.176, 58.073	24.176, 58.055
Average time ↓			10.433, 19.313	10.433, 19.309	11.381, 19.664	11.381, 19.659	9.649, 18.792	9.649, 18.788

(b) Training and test time for GP-pro models

Figure 5 shows the average RMSE, NLL, ENCE scores, and training time for GP-pro-c models using balltree, k -means, and random assignment methods with 100, 200, 300, 400, 500, and 600 data points per local GP, evaluated on four synthetic functions and four benchmark datasets (10d Ackley, 10d Levy, 10d Rastrigin, 50d Rosenbrock, 4d Power, 20d Parkinson, 8d Kinetic, and 9d Protein). RMSE and NLL decrease monotonically as the number of points per local GP increases, while training time increases accordingly. Models with 100 points per local GP exhibit the highest RMSE and NLL but the lowest training time, whereas models with 600 points per local GP achieve the best predictive accuracy at the cost of significantly higher computational overhead. Models with 200–400 points per local GP strike a favourable balance between predictive performance and computational cost.

Across different local GP sizes, it is observed that random assignment achieves lower absolute ENCE because weakly correlated local experts produce conservative predictive variances that tend to match empirical errors, despite degraded posterior mean accuracy. This results in higher RMSE and NLL but lower ENCE. In contrast, balltree and k -means preserve locality and yield more accurate posterior means, reducing RMSE and NLL, while inducing higher information redundancy across experts that makes uncertainty calibration more challenging and leads to higher ENCE.

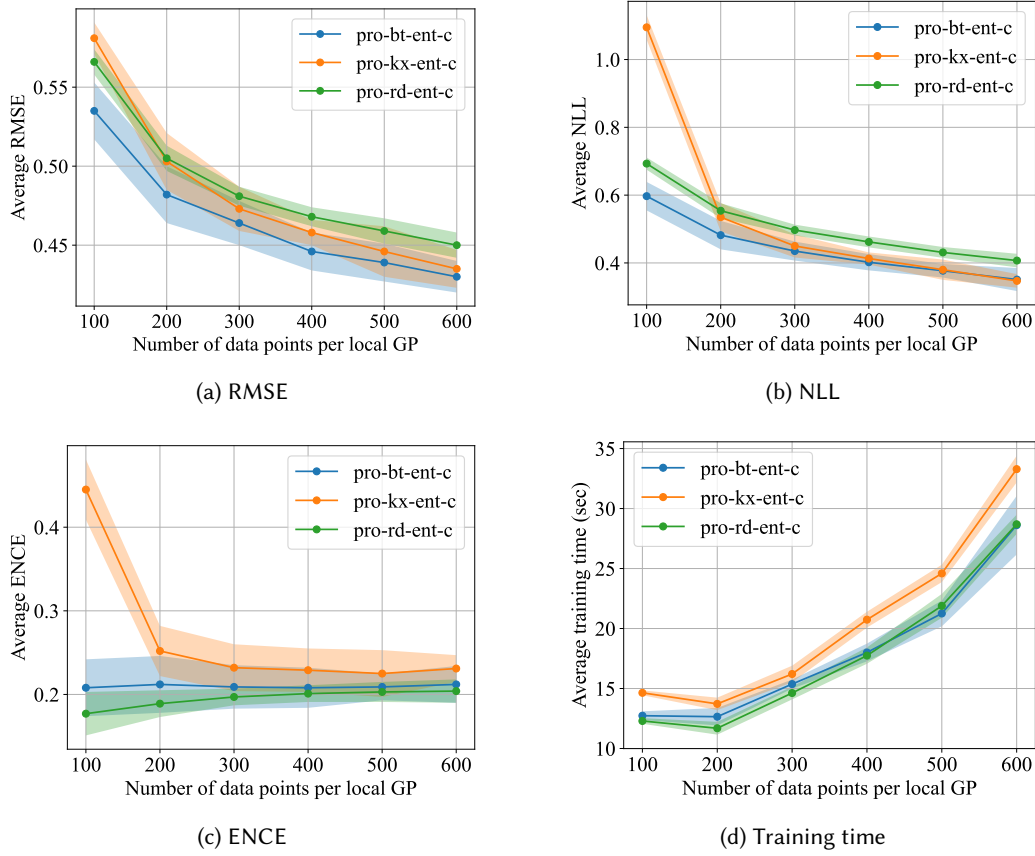


Fig. 5. Average NLL, RMSE, ENCE, and training time for GP-pro-c models using balltree, k -means, and random assignment methods with 100-600 data points per local GP, evaluated on four synthetic functions and four benchmark datasets. Random assignment yields lower ENCE but higher RMSE and NLL due to conservative, weakly correlated local GPs. Balltree and k -means preserve locality, yielding lower RMSE and NLL but higher ENCE. k -means performs poorly with small cluster sizes because small, high-dimensional clusters can be heterogeneous and poorly conditioned; performance improves sharply as cluster size increases. k -means can also incur slightly higher computational overhead due to uneven cluster sizes, whereas balltree and random assignments are more balanced. Overall, balltree assignment with 200-400 points per local GP provides a good trade-off between accuracy, calibration, and computational cost.

Notably, k -means performs poorly when the number of points per local GP is small, because small clusters in high-dimensional spaces can be heterogeneous and poorly conditioned. This leads to unstable local GP estimates and inflated NLL and ENCE, which improves sharply once each expert has sufficient data.

In terms of computational cost, k -means can be slightly more expensive than random or balltree assignment due to imbalanced cluster sizes. In contrast, balltree and random assignment tend to produce more balanced partitions, leading to more stable computational requirements.

Overall, balltree assignment with 200-400 data points per local GP provides the best trade-off between predictive accuracy, uncertainty calibration, and computational efficiency.

Table 9. Average RMSE, NLL, and ENCE values across four synthetic functions and six regression benchmark datasets, comparing GP-pro-c models with uncertainty calibration, GP-pro models without calibration, GP-pro-sm models using softmax-weight calibration, and a single global GP (GP-full). GP-pro-c models achieve lower NLL and ENCE scores, indicating that the proposed information-based calibration maintains predictive accuracy while improving uncertainty calibration.

Dataset	Size	Dim	GP-pro-bt-ent-c			GP-pro-bt-ent			GP-pro-sm			GP-full		
			RMSE↓	NLL↓	ENCE↓	RMSE↓	NLL↓	ENCE↓	RMSE↓	NLL↓	ENCE↓	RMSE↓	NLL↓	ENCE↓
Ackley	5000	10	0.487	0.742	0.190	0.487	0.775	0.245	0.527	0.801	0.115	0.468	0.663	0.057
Levy	5000	10	0.681	1.012	0.103	0.681	1.021	0.160	0.731	1.097	0.115	0.653	0.982	0.137
Rastrigin	5000	10	0.745	1.124	0.060	0.745	1.138	0.127	0.801	1.211	0.116	0.713	1.081	0.039
Rosenbrock	50000	50	0.610	0.918	0.116	0.610	0.929	0.167	0.711	1.106	0.162	N/A	N/A	N/A
Concrete	1030	8	0.336	0.287	0.162	0.336	0.287	0.166	0.341	0.298	0.186	0.335	0.282	0.190
Airfoil	1503	5	0.283	0.093	0.190	0.286	0.099	0.191	0.276	0.063	0.227	0.280	0.022	0.176
Parkinsons	5875	20	0.027	-1.374	0.750	0.027	-1.352	0.759	0.025	-1.387	0.761	0.016	-1.640	0.805
Power	9568	4	0.249	0.034	0.069	0.249	0.034	0.071	0.244	0.050	0.206	N/A	N/A	N/A
Kin40K	40000	8	0.384	0.466	0.306	0.375	0.471	0.348	0.229	-0.154	0.172	N/A	N/A	N/A
Protein	45730	9	0.664	0.932	0.097	0.663	0.930	0.086	0.741	1.462	0.725	N/A	N/A	N/A
Average ↓			0.447	0.423	0.204	0.446	0.433	0.232	0.463	0.455	0.278	N/A	N/A	N/A

6.3 Comparison with Baseline Methods

The third set of experiments compares the performance and computational cost of GP-pro-c against three baselines: (i) GP-pro without calibration, (ii) GP-pro with softmax weight calibration (Cohen et al. 2020), denoted GP-pro-sm, and (iii) a single global GP, denoted GP-full. All GP-pro and GP-pro-c variants use balltree assignment, entropy-based weighting, and 200 data points per local expert.

Table 9 reports performance metrics, Figure 6 shows reliability diagrams, and Table 10 summarises training and test times across four synthetic functions and six benchmark datasets.

Overall, GP-pro-c achieves lower RMSE, NLL, and ENCE than both GP-pro and GP-pro-sm. In addition, its reliability curves are consistently closer to the ideal diagonal, indicating improved uncertainty calibration. These results demonstrate that the information-based calibration method is doing a good job in maintaining accuracy performance while yielding better uncertainty calibration compared to the softmax weight calibration used in the GP-pro-sm models.

It is important to note that the softmax calibration method used in the GP-pro-sm model calibrates the aggregation weights rather than the uncertainty quantification of individual local experts, and relies on a global temperature parameter to control the degree of calibration. In our experiments, the temperature was fixed at 100, following Cohen et al. (2020). With this setting, GP-pro-sm achieves strong performance on datasets such as 8d Kin40K, where it outperforms both the uncalibrated GP-pro and the proposed GP-pro-c models in terms of RMSE, NLL and ENCE. However, on other datasets such as 9d Protein, the same temperature leads to degraded performance, with GP-pro-sm exhibiting higher RMSE, NLL and ENCE than GP-pro. GP-pro-sm behaves very differently on Kin40K and Protein because softmax calibration amplifies differences among local experts. Kin40K benefits from aggressive reweighting due to smoothness and expert agreement, whereas Protein suffers because expert heterogeneity and noise cause unreliable experts to dominate when the same temperature is used. These observations indicate that the effectiveness of softmax-based calibration is highly dataset-dependent and sensitive to the choice of temperature. In contrast, the proposed information-based calibration method provides more consistent performance across datasets without requiring dataset-specific hyperparameter tuning.

While Cohen et al. (2020) use a fixed temperature for softmax calibration, they do not propose automatic tuning; in principle, the temperature could be chosen via validation metrics such as negative log-likelihood or ENCE, but this adds additional hyperparameter optimisation and computational cost. Moreover, a single global temperature may be insufficient to capture the heterogeneous behaviour of local GP experts, particularly in

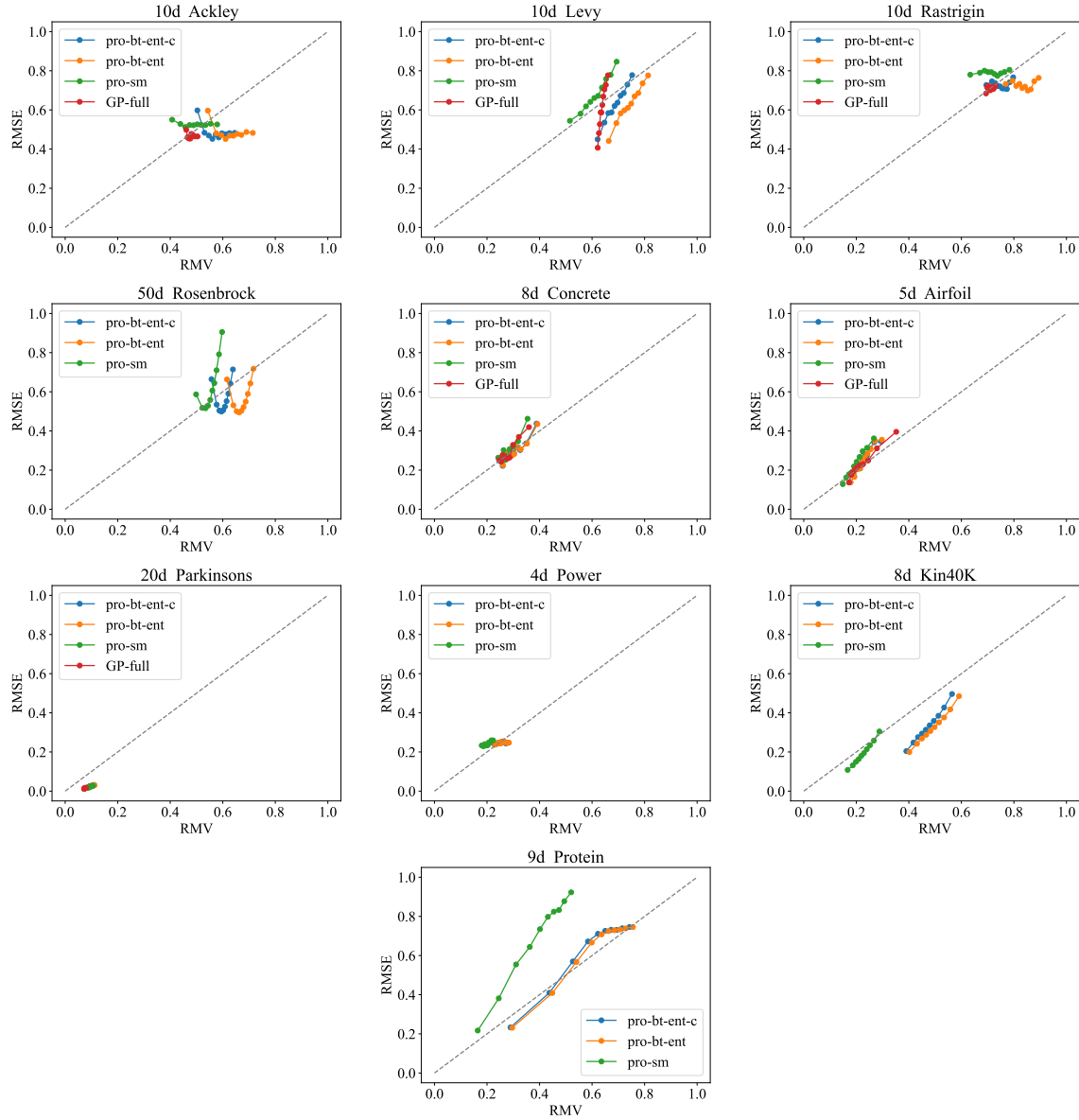


Fig. 6. Reliability diagrams for the GP-pro-c, GP-pro, GP-pro-sm, and GP-full models on four synthetic benchmark functions and six regression datasets. The GP-full models consistently exhibit curves closest to the diagonal. The GP-pro-c models demonstrate improved calibration compared to GP-pro across most datasets. On the Kin40K dataset, the GP-pro-sm model exhibits a curve closest to the diagonal. However, on the Protein dataset, GP-pro-sm exhibits noticeable deviation from the diagonal, whereas GP-pro-c maintains more stable calibration behaviour across datasets.

Table 10. Average training and test times (in seconds) across four synthetic functions and six regression benchmark datasets for GP-pro-c models with uncertainty calibration, GP-pro models without calibration, GP-pro-sm models using softmax-weight calibration, and a single global GP (GP-full).

Dataset	Size	Dim	GP-pro-bt-ent-c	GP-pro-bt-ent	GP-pro-sm	GP-full
			training, test time	training, test time	training, test time	training, test time
Ackley	5000	10	3.101, 1.427	3.101, 1.425	3.101, 1.425	47.905, 0.885
Levy	5000	10	3.380, 1.452	3.380, 1.450	3.380, 1.450	46.986, 0.892
Rastrigin	5000	10	3.119, 1.454	3.119, 1.452	3.119, 1.453	50.469, 1.053
Rosenbrock	50000	50	32.522, 85.483	32.522, 85.375	32.522, 85.695	N/A
Concrete	1030	8	1.263, 0.037	1.263, 0.036	1.263, 0.037	1.717, 0.033
Airfoil	1503	5	1.359, 0.066	1.359, 0.065	1.359, 0.066	2.870, 0.072
Parkinsons	5875	20	4.282, 2.581	4.282, 2.579	4.282, 2.580	77.394, 1.595
Power	9568	4	5.779, 0.538	5.779, 0.533	5.779, 0.532	N/A
Kin40K	40000	8	27.192, 58.806	27.192, 58.716	27.192, 59.187	N/A
Protein	45730	9	27.430, 59.042	27.430, 58.952	27.430, 59.441	N/A
Average ↓			10.433, 19.342	10.433, 19.313	10.433, 19.410	N/A

high-dimensional or large-scale settings. In contrast, GP-pro-c provides a robust, hyperparameter-free calibration of uncertainties.

In terms of computational cost, GP-pro-c and GP-pro-sm exhibit similar overhead. In contrast, the cost of GP-full increases rapidly with dataset size and dimensionality due to its cubic complexity. As a result, GP-full is omitted from experiments on datasets with more than 5,000 training points due to prohibitive runtime and memory requirements.

7 Discussion and Conclusions

This work addresses the overestimation of predictive uncertainty in product-of-experts Gaussian process (GP-pro) models, which arises from training local experts on disjoint data subsets. The proposed information-based variance calibration method is therefore tailored to GP-pro models. While this focus limits immediate applicability to a specific class of GP architectures, it enables us to exploit structural properties unique to the product-of-experts formulation, most notably, the discrepancy between the information available to individual experts and that available to the full dataset. This discrepancy admits a principled characterisation in terms of conditional mutual information, which yields a theoretically grounded calibration ratio for reducing overestimated posterior variances.

The proposed calibration preserves the computational advantages and aggregation structure of GP-pro while improving uncertainty estimates and maintaining predictive accuracy. Empirical results show consistent reductions in NLL and ENCE across synthetic functions and real-world regression datasets. On average, GP-pro-c reduces NLL and ENCE by 2.3% and 12.0%, respectively, compared to GP-pro without calibration. In terms of computational efficiency, GP-pro-c mitigates the cubic complexity of standard GP regression by operating on local experts. For datasets where a global GP (GP-glo) is tractable, GP-pro-c achieves an average reduction of 72.5% in computational cost. For larger datasets, GP-glo becomes infeasible due to runtime or memory constraints, whereas GP-pro-c remains tractable, highlighting its practical scalability.

Among the GP-pro-c variants, the combination of balltree assignment and entropy-based weighting performs best overall. Empirical results further suggest that using 200–400 data points per expert provides a favourable balance between predictive accuracy, calibration quality, and computational cost.

The results presented in this study demonstrate consistent, if moderate, improvements in expected normalised calibration error (ENCE) for the GP-pro-c model compared to the baseline GP-pro. This is expected, as the proposed method applies a controlled correction based on information gain rather than aggressively reshaping

the posterior distribution. Importantly, these improvements are achieved without degrading predictive accuracy, as reflected by stable RMSE values across all experiments. Furthermore, it is important to note that ENCE is inherently bounded below by model misspecification and observational noise. Reliability diagrams support these findings: although GP-pro-c does not uniformly outperform GP-pro across all probability levels, it consistently moves predictive uncertainty closer to ideal calibration in regions where GP-pro exhibits inflated variance. Overall, the method improves uncertainty calibration in a conservative and robust manner.

Several directions for future work remain. First, extending the method to alternative kernel classes is a promising avenue. Although the derivation relies on standard assumptions related to normalised information gain (with reference to Matérn kernels), the calibration rule itself depends only on posterior variances and noise levels. This suggests that the approach may generalise to other stationary kernels, although formal analysis is left for future work. Second, incorporating joint calibration across experts and explicitly modelling dependencies between local GPs may further reduce variance overestimation, particularly in overlapping regions. Third, establishing theoretical guarantees for posterior variance calibration would strengthen the methodological foundation of GP-pro-c.

Another promising direction is the integration of GP-pro-c with dimensionality-reduction approaches, such as principal component-based methods (Higdon et al. 2008). These approaches are complementary: GP-pro-c addresses scalability and uncertainty calibration in large input spaces, while dimensionality reduction targets high-dimensional outputs. Combining these strategies could enable efficient modelling of problems with both large input datasets and high-dimensional outputs.

Finally, integrating GP-pro-c into Bayesian optimisation (BO) pipelines represents a natural direction for future work. Gaussian processes are widely used as surrogate models in BO (Brochu et al. 2010; Frazier 2018; Garnett 2023; Shahriari et al. 2016), yet their scalability remains a major challenge. Moreover, BO is particularly sensitive to the quality of uncertainty estimates, as posterior variances directly determine the behaviour of acquisition functions. Overestimated uncertainties can encourage excessive exploration, leading to less efficient optimisation. By combining scalability with improved uncertainty calibration, GP-pro-c is well positioned to address these challenges, potentially enabling more reliable acquisition decisions and greater optimisation efficiency. Although a comprehensive evaluation within BO is beyond the scope of the present paper, the calibration improvements demonstrated here provide a strong foundation for future investigation of GP-pro-c as a scalable and reliable surrogate model for BO.

Acknowledgments

The first author gratefully acknowledges Rasul Tutunov, Haitham Bou-Ammar, and Sye Loong Keoh for their valuable feedback and insightful discussions during the early development of the information-based calibration method. The authors also thank the anonymous reviewers for their constructive comments and suggestions, which have helped improve the clarity and quality of this paper.

References

- D. Arthur and S. Vassilvitskii. 2007. “K-Means++: The Advantages of Careful Seeding.” In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA '07). Society for Industrial and Applied Mathematics, USA, 1027–1035.
- E. Brochu, V. M. Cora, and N. de Freitas. 2010. *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. arXiv: [1012.2599](https://arxiv.org/abs/1012.2599) (cs.LG).
- Y. Cao. 2018. “Scaling Gaussian Processes.” Ph.D. Dissertation. University of Toronto.
- Y. Cao and D. J. Fleet. 2014. *Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions*. arXiv: [1410.7827](https://arxiv.org/abs/1410.7827) (cs.LG).
- Y. Cao and D. J. Fleet. 2015. *Transductive Log Opinion Pool of Gaussian Process Experts*. arXiv: [1511.07551](https://arxiv.org/abs/1511.07551) (cs.LG).
- K. Chaloner and I. Verdinelli. 1995. “Bayesian Experimental Design: A Review.” *Statistical Science*, 10, 3, 273–304.

- S. Cohen, R. Mubvha, T. Marwala, and M. P. Deisenroth. 2020. "Healing Products of Gaussian Process Experts." In: *Proceedings of the 37th International Conference on Machine Learning* (Proceedings of Machine Learning Research) Article 193. Vol. 119. JMLR.org.
- T. M. Cover and J. A. Thomas. 2005. "Entropy, Relative Entropy, and Mutual Information." In: *Elements of Information Theory*. John Wiley and Sons, Ltd. Chap. 2, 13–55. doi:[10.1002/047174882X.ch2](https://doi.org/10.1002/047174882X.ch2).
- M. Deisenroth and J. W. Ng. 2015. "Distributed Gaussian Processes." In: *Proceedings of the 32nd International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by F. Bach and D. Blei. Vol. 37. PMLR, Lille, France, 1481–1490.
- T. Desautels, A. Krause, and J. W. Burdick. 2014. "Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization." *Journal of Machine Learning Research*, 15, 1, 3873–3923.
- L. Dorard. 2012. "Bandit Algorithms for Searching Large Spaces." Ph.D. Dissertation. University College London.
- P. I. Frazier. 2018. *A Tutorial on Bayesian Optimization*. arXiv: [1807.02811](https://arxiv.org/abs/1807.02811) (stat.ML).
- J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. 2018. "GPpyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration." In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 7587–7597.
- R. Garnett. 2023. *Bayesian Optimization*. Cambridge University Press.
- T. Heskes. 1997. "Selecting Weighting Factors in Logarithmic Opinion Pools." In: *Advances in Neural Information Processing Systems*. Ed. by M. Jordan, M. Kearns, and S. Solla. Vol. 10. MIT Press.
- D. Higdon, J. Gattiker, B. Williams, and M. Rightley. 2008. "Computer Model Calibration Using High-Dimensional Output." *Journal of the American Statistical Association*, 103, 482, 570–583. doi:[10.1198/016214507000000888](https://doi.org/10.1198/016214507000000888).
- C.-W. Ko, J. Lee, and M. Queyranne. 1995. "An Exact Algorithm for Maximum Entropy Sampling." *Operations Research*, 43, 4, 684–691. doi:[10.1287/opre.43.4.684](https://doi.org/10.1287/opre.43.4.684).
- A. Krause, A. Singh, and C. Guestrin. 2008. "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies." *Journal of Machine Learning Research*, 9, 235–284.
- D. Levi, L. Gispán, N. Giladi, and E. Fetaya. 2020. *Evaluating and Calibrating Uncertainty Prediction in Regression Tasks*. arXiv: [1905.11659](https://arxiv.org/abs/1905.11659) (cs.LG).
- H. Liu, J. Cai, Y.-S. Ong, and Y. Wang. 2019. "Understanding and Comparing Scalable Gaussian Process Regression for Big Data." *Knowledge-Based Systems*, 164, 324–335. doi:[10.1016/j.knosys.2018.11.002](https://doi.org/10.1016/j.knosys.2018.11.002).
- H. Liu, Y.-S. Ong, X. Shen, and J. Cai. 2020. "When Gaussian Process Meets Big Data: A Review of Scalable GPs." *IEEE Transactions on Neural Networks and Learning Systems*, 31, 11, 4405–4423. doi:[10.1109/TNNLS.2019.2957109](https://doi.org/10.1109/TNNLS.2019.2957109).
- Z. Liu, L. Zhou, H. Leung, and H. P. H. Shum. 2016. "Kinect Posture Reconstruction Based on a Local Mixture of Gaussian Process Models." *IEEE Transactions on Visualization and Computer Graphics*, 22, 11, 2437–2450.
- G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. 1978. "An Analysis of Approximations for Maximizing Submodular Set Functions—I." *Mathematical Programming*, 14, 1, 265–294. doi:[10.1007/BF01588971](https://doi.org/10.1007/BF01588971).
- T. Nguyen and E. Bonilla. 2014. "Fast Allocation of Gaussian Process Experts." In: *Proceedings of the 31st International Conference on Machine Learning* (Proceedings of Machine Learning Research). Vol. 32. PMLR, 145–153.
- S. M. Omohundro. 1989. *Five Balltree Construction Algorithms*. Tech. rep. TR-89-063. International Computer Science Institute.
- C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.
- B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. 2016. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proceedings of the IEEE*, 104, 1, 148–175.
- N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. 2010. "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design." In: *Proceedings of the 27th International Conference on Machine Learning*. Omnipress, 1015–1022.
- N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. 2012. "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting." *IEEE Transactions on Information Theory*, 58, 5, 3250–3265. doi:[10.1109/TIT.2011.2182033](https://doi.org/10.1109/TIT.2011.2182033).
- B. Szabó and H. van Zanten. 2019. "An Asymptotic Analysis of Distributed Nonparametric Methods." *Journal of Machine Learning Research*, 20, 87, 1–30.
- K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi. 2020. "Methods for Comparing Uncertainty Quantifications for Material Property Predictions." *Machine Learning: Science and Technology*, 1, 2, 025006. doi:[10.1088/2632-2153/ab7e1a](https://doi.org/10.1088/2632-2153/ab7e1a).
- V. Tresp. 2000. "A Bayesian Committee Machine." *Neural Computation*, 12, 11, 2719–2741. doi:[10.1162/089976600300014908](https://doi.org/10.1162/089976600300014908).

A Information Gain

Equation (14) shows that the information gain for the data points in the set $\mathcal{A} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ can be expressed in terms of predictive variances. This result was established by Srinivas et al. (2012) and Dorard (2012) as follows.

In Equation (12), the first entropy term, $H(\mathbf{y}_{\mathcal{A}})$ or equivalently $H(\mathbf{y}_N)$, can be expressed recursively as $H(\mathbf{y}_N) = H(\mathbf{y}_{N-1}) + H(\mathbf{y}_N | \mathbf{y}_{N-1})$. Conditioned on $\mathbf{y}_{N-1}$, the inputs $(\mathbf{x}_n)_{1 \leq n \leq N}$ are deterministic, and thus $f(\mathbf{x}) \sim \mathcal{N}(0, \sigma_{N-1}^2(\mathbf{x}))$ for all $\mathbf{x}$.

Here, note that $\mathbf{x}_1, \dots, \mathbf{x}_N$ are deterministic given $\mathbf{y}_{N-1}$, and that the conditional variance $\sigma_{N-1}^2(\mathbf{x}_N)$ does not depend on $\mathbf{y}_{N-1}$. Meanwhile, $y_N = f(\mathbf{x}_N) + \epsilon_N$ is the sum of two independent zero-mean Gaussian random variables: one with variance $\sigma_{N-1}^2(\mathbf{x}_N)$ and the other with variance $\check{\sigma}^2$ (the observation noise variance).

By applying the entropy formula $H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \log |2\pi e \Sigma|$, we obtain

$$H(y_N | \mathbf{y}_{N-1}) = \frac{1}{2} \log (2\pi e (\sigma_{N-1}^2(\mathbf{x}_N) + \check{\sigma}^2))$$

Thus, expressing $H(\mathbf{y}_N)$ recursively yields

$$H(\mathbf{y}_N) = \frac{1}{2} \sum_{n=1}^N \log (2\pi e (\check{\sigma}^2 + \sigma_{n-1}^2(\mathbf{x}_n)))$$

For the second term in Equation (12), namely $H(\mathbf{y}_{\mathcal{A}} | \mathbf{f})$ or equivalently $H(\mathbf{y}_N | \mathbf{f})$, note that conditioned on $\mathbf{f}$, the vector $\mathbf{y}_N$ follows a zero-mean Gaussian distribution with covariance matrix $\check{\sigma}^2 \mathbf{I}_N$. Therefore,

$$\begin{aligned} H(\mathbf{y}_N | \mathbf{f}) &= \frac{1}{2} \log |2\pi e \check{\sigma}^2 \mathbf{I}_N| \\ &= \frac{1}{2} \sum_{n=1}^N \log (2\pi e \check{\sigma}^2) \end{aligned}$$

Combining these two terms, we obtain

$$\begin{aligned} I(\mathbf{y}_{\mathcal{A}}; \mathbf{f}) &= H(\mathbf{y}_N) - H(\mathbf{y}_N | \mathbf{f}) \\ &= \frac{1}{2} \sum_{n=1}^N \log (2\pi e (\check{\sigma}^2 + \sigma_{n-1}^2(\mathbf{x}_n))) - \frac{1}{2} \sum_{n=1}^N \log (2\pi e \check{\sigma}^2) \\ &= \frac{1}{2} \sum_{n=1}^N \log \left(\frac{2\pi e \check{\sigma}^2 + 2\pi e \sigma_{n-1}^2(\mathbf{x}_n)}{2\pi e \check{\sigma}^2} \right) \\ &= \frac{1}{2} \sum_{n=1}^N \log \left(1 + \frac{\sigma_{n-1}^2(\mathbf{x}_n)}{\check{\sigma}^2} \right) \end{aligned}$$

Here, $\sigma_{n-1}^2(\mathbf{x}_n)$ denotes the posterior variance of $f(\mathbf{x}_n)$ given observations $\{y_1, \dots, y_{n-1}\} \subseteq \mathbf{y}_{\mathcal{A}}$. As given in Equation (3), this quantity does not depend on the observed values themselves.

Since the sum can therefore be evaluated without observing the data, it is possible to compute the mutual information that would be gained from any hypothetical set of observations (Desautels et al. 2014; Srinivas et al. 2012).

B Additional Experimental Results

Figure 7 shows reliability diagrams for the GP-pro-c and GP-pro models using the balltree assignment method, comparing entropy-, variance-, and uniform-weighted schemes. It is observed that, under all weighting schemes, variants of the GP-pro-c model produce curves closer to the ideal diagonal line than the corresponding GP-pro models.

Received 10 September 2025; accepted 15 June 2026

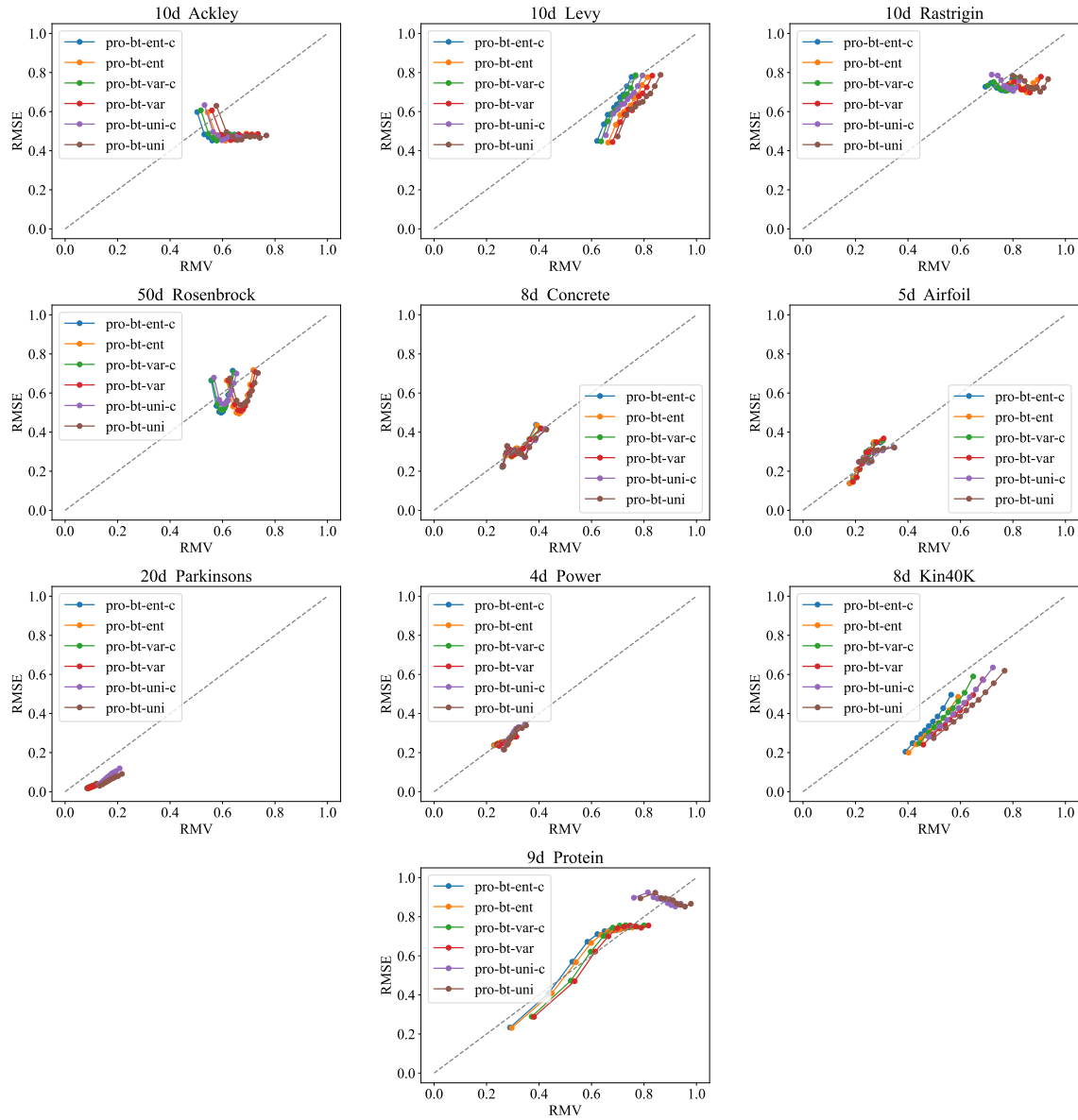


Fig. 7. Reliability diagrams for the GP-pro-c and GP-pro models using the balltree assignment method, comparing entropy-, variance-, and uniform-weighted schemes on four synthetic benchmark functions and six regression datasets. Under all weighting schemes, GP-pro-c models with uncertainty calibration produce curves closer to the ideal diagonal line than GP-pro models without uncertainty calibration.