HMTMO-GP: Hierarchical Multi-Task Multi-Output Gaussian Processes

Anonymous Authors[1]

Abstract

Machine learning models often employ a multi-task/multi-output (MTMO) architecture to tackle data scarcity by exploiting structural correlation. In this work, we extend Gaussian processes (GP) for hierarchical MTMO problems, referred to as HMTMO-GP, which leverages GP probabilistic modeling for data-efficient learning and intrinsic uncertainty estimation. Our results on gene expression time series and an antibody solution viscosity dataset demonstrate that HMTMO-GP delivers strong predictive accuracy and well-calibrated uncertainty estimates compared to existing multi-task or multi-output GP models.

1. Introduction

Pharmaceutical datasets are often hierarchical and may contain fundamental explainable patterns or shared structure across related studies, such as molecules with similar properties or experiments built around related process steps. As a result, measurements across studies are not independent. Multi-task/multi-output (MTMO) learning frameworks (Journel & Huijbregts, 1978; Xu et al., 2020; Yu et al., 2024) can exploit the task/output correlations to improve data efficiency. In this work, MTMO distinguishes between tasks and outputs, as our use cases involve multiple hierarchically structured tasks, each associated with multiple correlated outputs; we note that the terms multi-task (MT) and multi-output (MO) are often used interchangeably in the literature. MTMO approaches can thus be effective means to reduce time and cost in data-scarce pharmaceutical workflows.

Several probabilistic MTMO models have been proposed to enable principled uncertainty quantification, including process convolution models (Higdon, 2002), clustered multi-task Gaussian processes (Rasmussen & Ghahramani, 2002), Gaussian process regression networks (Wilson et al., 2012), multi-output spectral mixture Gaussian processes (Parra & Tobar, 2017), deep Gaussian processes (Salimbeni & Deisenroth, 2017), and more recent graphical or deep hierarchical extensions (Dai et al., 2024; Chang & Sung, 2025). These approaches differ in how they parameterize correlations, yet jointly they offer limited mechanisms for explicitly controlling how information propagates across hierarchical task structures. To address this gap, we propose *HMTMO-GP*, a hierarchical MTMO Gaussian process framework that captures heterogeneous correlations by decomposing the covariance into level-specific components with learnable weights, enabling structured and controllable information sharing that better matches hierarchical data.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Methods

2.1. Overview of Gaussian Process (GP)

Throughout this paper, we assume a shared input space $X = x_1, \ldots, x_N \subset \mathbb{R}^d$. A GP defines a joint Gaussian distribution over a (noise-free) function $f(x)$ through a mean function $m(x)$ and a covariance kernel $\mathrm{Cov}[f(x), f(x')] = k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \,. \tag{1}$$

Conditioning on observed data $(x, f(x))$ and new inputs x^* yields closed-form posterior predictions $f(x^*)$, making GPs well suited for data-efficient forecasting.

2.2. Multi-output Gaussian Process Models

Multi-output GP models extend standard GPs by defining structured covariance functions over the outputs. In what follows, the output is indexed by $p \in \{1, \ldots, P\}$. The main distinction among these methods lies in how covariance is defined and shared across outputs, for instance, additive or multiplicative kernels or their linear combinations. We briefly review **Linear Model of Coregionalization (LMC)** (Journel & Huijbregts, 1978; Álvarez & Lawrence, 2011) and **Hierarchical Gaussian process (HGP)** (Hensman et al., 2013; Ma et al., 2023), which provide the two ingredients integrated in our *HMTMO-GP*: the kernel-mixture flexibility of LMC and the hierarchical inheritance structure of HGP.

Let $\mathbf{f}(x) = [f_1(x), \ldots, f_P(x)]^\top$ denote the vector of outputs at input x. In **LMC**, each output (we consider the

noise-free scenario for notational clarity) is expressed as a linear combination of Q shared latent processes $u_q(x)$,

$$f_p(x) = \sum_{q=1}^{Q} a_{p,q}\, u_q(x), \quad u_q(\cdot) \sim \mathcal{GP}(0, k_q(\cdot,\cdot)). \quad (2)$$

Here, $a_{p,q}$ is the mixing weight that determines how strongly latent process q contributes to output p. Defining $\mathbf{a}^{(q)} = [a_{1,q}, \ldots, a_{P,q}]^\top$, the covariance weight induced by latent process $u_q(x)$ is $B^{(q)} = \mathbf{a}^{(q)}\mathbf{a}^{(q)\top}$, leading to

$$\mathrm{Cov}[\mathbf{f}(x), \mathbf{f}(x')] = \sum_{q=1}^{Q} B^{(q)}\, k_q(x, x'). \quad (3)$$

LMC encodes multi-output dependencies through a uniform linear mixing shared across all tasks and inputs. While computationally efficient and interpretable, this unconditional mixing can be restrictive when correlation strengths vary across outputs.

HGP models address this limitation by introducing MTMO dependence through a parent–child task hierarchy derived from the structure of the underlying dataset. In the standard formulation, the output function of each child task inherits from a shared parent function,

$$g(x) \sim \mathcal{GP}(0, k_g), \quad f_p(x) \mid g(x) \sim \mathcal{GP}(g(x), k_f), \quad (4)$$

such that task-specific output functions take the parent process as the mean function while retaining their own variation through k_f. As in most MTMO models, the set of outputs is fixed across all hierarchy levels. In HGP, however, inheritance occurs only at the function level through a single shared parent function. Consequently, all tasks share the same hierarchical structure, which limits explicit control over how covariance relationships vary across task groups or hierarchy levels.

2.3. Hierarchical Multi-Task Multi-Output Gaussian Processes (HMTMO-GP)

Motivated by the complementary strengths of LMC and HGP, we propose HMTMO-GP, which combines the kernel-mixture flexibility of LMC with the hierarchical inheritance structure of HGP. Latent processes evolve recursively across hierarchy levels, with inheritance coefficients controlling the strength of information propagation and kernel mixing weights determining how latent processes contribute to task outputs. At level $\ell \in \{1, \ldots, L\}$, for input x, the hypothetical signal is modeled as

$$z_p^{(\ell)}(x) = \sum_{q=1}^{Q_\ell} a_{p,q}^{(\ell)} u_q^{(\ell)}(x), \quad (5)$$

where $a_{p,q}^{(\ell)}$ are kernel mixing weights (as in LMC), and $\mathbf{u}^{(\ell)}(x) = \left[u_1^{(\ell)}(x), \ldots, u_{Q_\ell}^{(\ell)}(x)\right]^\top$ collects the latent processes at level ℓ.

At the root level, $z^{(0)}(x) = u^{(0)}(x)$ with

$$u^{(0)}(x) \sim \mathcal{GP}\Big(0, k^{(0)}(x, x')\Big), \quad (6)$$

For $\ell \geq 1$, each latent process is conditionally generated from the inherited signal

$$u_q^{(\ell)}(x) \mid z^{(\ell-1)}(x) \sim \mathcal{GP}\Big(\rho_q^{(\ell)} z^{(\ell-1)}(x),\, k_q^{(\ell)}(x, x')\Big), \quad (7)$$

where $\rho_q^{(\ell)}$ governs the strength of hierarchical inheritance.

Unrolling this recursion yields the explicit decomposition

$$u_q^{(\ell)}(x) = \alpha_q^{(\ell)}\, u^{(0)}(x) + \sum_{s=1}^{\ell} \alpha_q^{(\ell,s)}\, \epsilon_q^{(s)}(x), \quad (8)$$

where $\alpha_q^{(\ell)} = \prod_{r=1}^{\ell} \rho_q^{(r)}$, $\alpha_q^{(\ell,s)} = \prod_{j=s+1}^{\ell} \rho_q^{(j)}$, and $\epsilon_q^{(s)}(\cdot) \sim \mathcal{GP}\Big(0, k_q^{(s)}(\cdot,\cdot)\Big)$ are independent innovation processes.

Substituting Eq. (8) into Eq. (5), the leaf-level output at $\ell = L$ becomes

$$f_p^{(L)}(x) = \sum_{q=1}^{Q_L} a_{p,q}^{(L)} \left[\alpha_q^{(L)}\, u^{(0)}(x) + \sum_{s=1}^{L} \alpha_q^{(L,s)}\, \epsilon_q^{(s)}(x)\right]. \quad (9)$$

Assuming independence across latent processes and innovations, the leaf-level covariance is

$$\mathrm{Cov}\Big[f_p^{(L)}(x), f_{p'}^{(L)}(x')\Big] = \sum_{q=1}^{Q_L} a_{p,q}^{(L)} a_{p',q}^{(L)} \sum_{s=0}^{L} \Gamma_{q,s}\, k_q^{(s)}(x, x'),$$
$$\Gamma_{q,0} = \prod\nolimits_{r=1}^{L} \rho_q^{(r)2},\ \Gamma_{q,s} = \prod\nolimits_{j=s+1}^{L} \rho_q^{(j)2}, \quad (10)$$

where $s = 1, \ldots, L$. Stacking all outputs across training inputs into $\mathbf{f} \in \mathbb{R}^{|\mathcal{P}|N}$ yields the joint covariance

$$\mathbf{K} = \sum_{\ell=0}^{L} \sum_{q=1}^{Q_\ell} \mathbf{B}_q^{(\ell)} \otimes \mathbf{K}_{q,X}^{(\ell)}, \quad (11)$$

where $(\mathbf{K}_{q,X}^{(\ell)})_{n,n'} = k_q^{(\ell)}(x_n, x_{n'})$, $\mathbf{B}_{q,p,p'}^{(\ell)} = a_{p,q}^{(\ell)} a_{p',q}^{(\ell)}$.

Given training observations $\mathbf{y}$ and test inputs $\mathbf{X}_*$, the posterior predictive distribution is

$$p(\mathbf{f}_* \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_*) = \mathcal{N}\big(\mathbf{K}_*\mathbf{K}^{-1}\mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_*\mathbf{K}^{-1}\mathbf{K}_*^\top\big), \quad (12)$$

where $\mathbf{K}_*$ denotes the cross-covariance between training and test inputs and $\mathbf{K}_{**}$ is the covariance at the test inputs.

3. Results

We now evaluate HMTMO-GP empirically on hierarchical forecasting benchmarks to assess its predictive accuracy and uncertainty calibration relative to LMC and HGP under heterogeneous task structures.

3.1. Gene Expression Dataset

The first benchmark is a gene expression time-series dataset (Hensman et al., 2013; Ma et al., 2023), in which the MTMO setting involves predicting microarray mRNA expression intensities over up to 13 time points for 3,590 genes measured across eight biological replicates, with time as the sole input variable. To capture structured correlations among genes, we construct (Hensman et al., 2014) a top-level hierarchy by clustering genes based on latent temporal trajectory similarity. This hierarchy provides a principled setting for evaluating how models balance information sharing across clusters with gene- and replicate-specific variation.

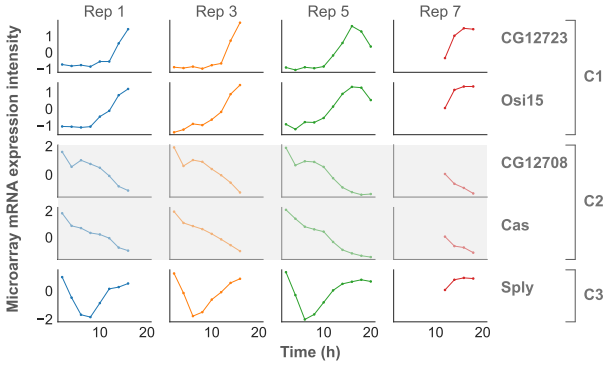

Figure 1. Gene expressions for selected genes and replicates in the three-cluster (C1–C3) evaluation. Each row corresponds to a gene and each column to a biological replicate (Rep 1, 3, 5, and 7).

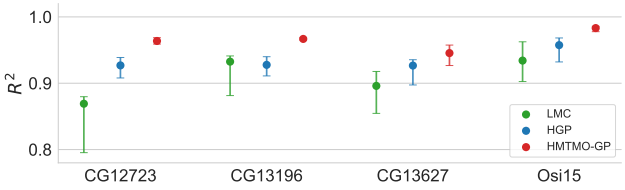

Figure 2. R^2 performance for four genes drawn from the same cluster. All models benefit from strong coherence, with HMTMO-GP achieving the highest accuracy and lowest variability.

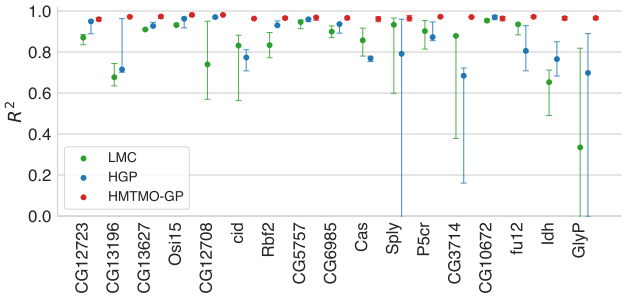

Figure 3. R^2 performance for genes drawn from three clusters with heterogeneous expression dynamics. HMTMO-GP achieves the highest accuracy and greatest stability across genes.

Figure 1 illustrates a subset of the genes from these clusters, highlighting the hierarchical structure used in the evaluation.

To assess model robustness under increasing heterogeneity, we consider two settings by selecting genes from one and three clusters, respectively (Figures 2–3). All models are evaluated using identical train–test splits with five-fold cross-validation. Performance is assessed using R^2 and mean negative log predictive density (MNLP), which assess point-wise accuracy and probabilistic calibration, respectively.

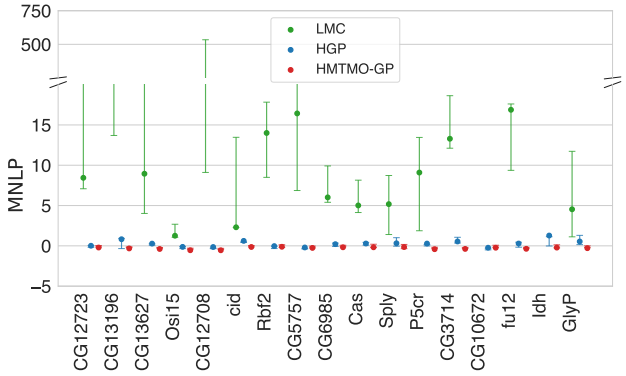

Figure 4. MNLP comparison for the three-cluster gene expression setting. The broken y-axis accommodates extreme MNLP values for LMC while preserving resolution for remaining methods. HMTMO-GP consistently achieves lower MNLP, indicating superior uncertainty calibration under heterogeneous dynamics.

Across both settings, HMTMO-GP attains the highest median R^2 with the smallest dispersion (Figure 3), indicating robust performance across genes with diverse temporal dynamics. In contrast, HGP performs competitively in the most coherent (single-cluster) setting (Figure 2), but becomes less stable as cluster diversity increases, while LMC consistently underperforms due to globally uniform task coupling. MNLP results (Figure 4) further show that HMTMO-GP provides better calibrated probabilistic forecasts under multi-cluster dynamics.

3.2. Antibody Residue Viscosity Dataset

We further evaluate model performance on an antibody residue viscosity dataset (Cloutier et al., 2020), which contains more than 40,000 viscosity measurements collected across residue-level variations in the antibody sequence, each measured under six excipient conditions. Figure 5 summarizes the task hierarchy, inputs and outputs of this dataset. To emulate the data scarcity commonly encountered in pharmaceutical applications, we subsample 1,500 residues for model evaluation. All models follow the same evaluation protocol as used for the gene expression dataset.

For Support Vector Machine (SVM) and Elastic Net (EN)

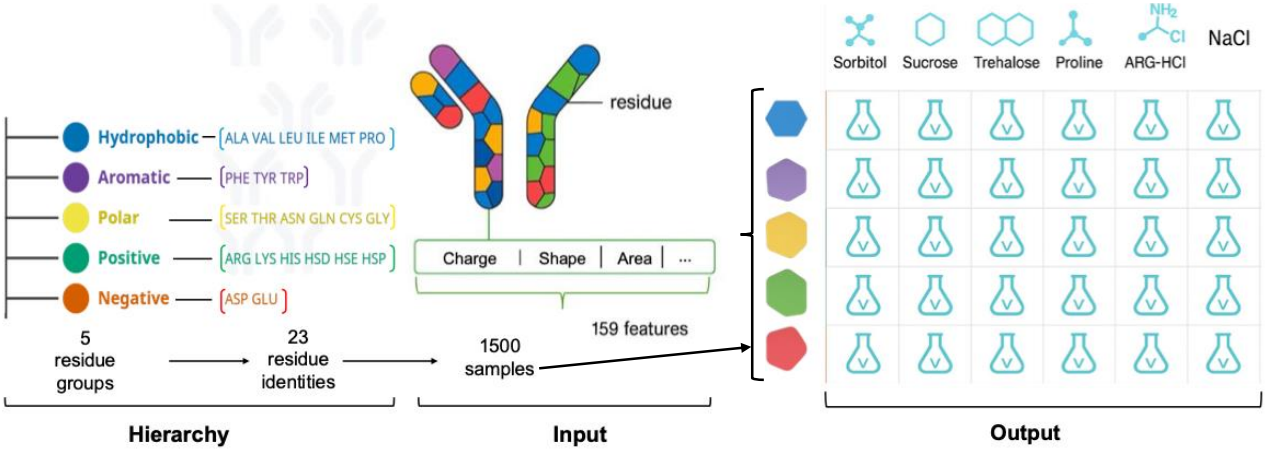

Figure 5. Hierarchy, inputs, and outputs of the antibody viscosity dataset. Antibodies are decomposed into a sequence of residues from 23 amino acid identities. Residue identities are grouped into five physicochemically motivated categories, inducing a hierarchy that supports structured information sharing. Each observation relating to a single residue variation is represented by 159 physicochemical and structural features. Formulation conditions are defined by six excipient identities (sorbitol, sucrose, trehalose, proline, arg·HCl, and NaCl).

baselines, we adopt the hyperparameters reported in the original study and evaluate them on the 1,500-point subset used here; the corresponding RMSE values are reported in Table 1.

Table 1. Average RMSE for viscosity predictions under different excipients. SVM and EN use hyperparameters from the original study and are evaluated on the 1,500-point subset used in this work.

Excipient	SVM	EN	HMTMO-GP
Sorbitol	0.0254	0.0258	**0.0215**
Sucrose	0.0338	0.0308	**0.0284**
Trehalose	0.0269	0.0262	**0.0238**
Proline	0.0258	0.0247	**0.0219**
Arg_HCl	0.1443	0.1259	**0.1173**
NaCl	0.0236	0.0269	**0.0201**
Mean	0.0476	0.0434	**0.0388**

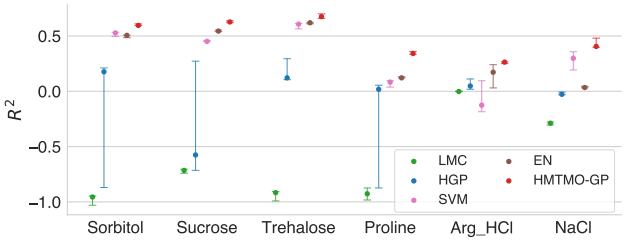

Figure 6. R^2 performance across six excipients on the antibody residue viscosity dataset. HMTMO-GP achieves higher predictive accuracy with lower variability than LMC, HGP, SVM, and EN.

As shown in Figures 6 and 7, non-hierarchical approaches exhibit inconsistent performance across excipients, while LMC shows large variability due to globally uniform task coupling. HGP improves accuracy for some outputs but remains sensitive to mismatches in task relatedness across hierarchy levels. Classical baselines (SVM and EN) yield more stable but generally weaker accuracy and calibration. In contrast, HMTMO-GP consistently attains higher R^2 and lower MNLP across all excipients, indicating both improved predictive performance and better-calibrated uncertainty through hierarchy-aware information sharing.

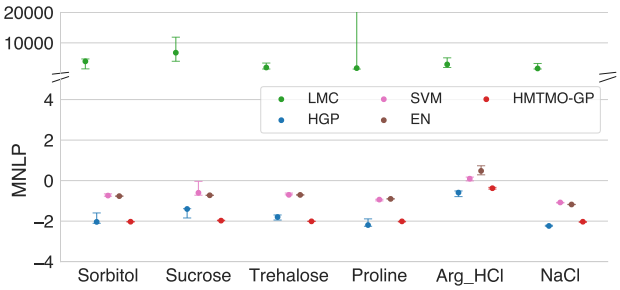

Figure 7. MNLP across excipients for all methods. HMTMO-GP consistently attains lower MNLP, indicating improved uncertainty calibration under hierarchical task structure.

4. Conclusions and Discussion

HMTMO-GP combines level-specific kernel mixtures with structured inheritance to model hierarchical data within a unified probabilistic framework. For pharmaceutical datasets such as gene expression and antibody viscosity—which are hierarchically organized and costly to generate—this approach improves data efficiency while achieving higher predictive accuracy and better-calibrated uncertainty by aligning covariance structure with the underlying data hierarchy. Evaluating the framework on datasets beyond pharmaceutical science would further assess its generality for accurately representing hierarchical data structures.

References

Chang, C.-Y. and Sung, C.-L. Deep intrinsic coregionalization multi-output gaussian process surrogate with active learning. *arXiv preprint arXiv:2508.16434*, 2025.

Cloutier, T. K., Sudrik, C., Mody, N., Sathish, H. A., and Trout, B. L. Machine learning models of antibody–excipient preferential interactions for use in computational formulation design. *Molecular Pharmaceutics*, 17 (9):3589–3599, 2020. doi: 10.1021/acs.molpharmaceut.0c00629.

Dai, Y., Yan, W., and Yin, F. Graphical multi-output gaussian process with attention. In *International Conference on Learning Representations*, 2024.

Hensman, J., Lawrence, N. D., and Rattray, M. Hierarchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, 14(1):252, 2013. doi: 10.1186/1471-2105-14-252.

Hensman, J., Rattray, M., and Lawrence, N. D. Fast nonparametric clustering of structured time-series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. doi: 10.1109/TPAMI.2014.2318711.

Higdon, D. Space and space-time modeling using process convolutions. *Quantitative Methods for Current Environmental Issues*, pp. 37–56, 2002.

Journel, A. and Huijbregts, C. *Mining Geostatistics*. Academic Press, 1978.

Ma, C., Leroy, A., and Álvarez, M. Latent variable multiouput gaussian processes for hierarchical datasets. *arXiv preprint arXiv:2308.16822*, 2023.

Parra, D. and Tobar, F. Spectral mixture kernels for multivariate gaussian processes. *Advances in Neural Information Processing Systems*, 2017.

Rasmussen, C. E. and Ghahramani, Z. Infinite mixtures of gaussian process experts. *Advances in Neural Information Processing Systems*, 2002.

Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for deep gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2421–2430, 2017.

Wilson, A. G., Knowles, D. A., and Ghahramani, Z. Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1139–1146, Edinburgh, Scotland, 2012.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C., and Shen, X. A survey on multi-output learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):546–561, 2020. doi: 10.1109/TNNLS.2019.2945133.

Yu, J., Dai, Y., Liu, X., Huang, J., Shen, Y., Zhang, K., Zhou, R., Adhikarla, E., Ye, W., Liu, Y., Kong, Z., Zhang, K., Yin, Y., Namboodiri, V., Davison, B. D., Moore, J. H., and Chen, Y. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. *arXiv preprint arXiv:2404.18961*, 2024.

Álvarez, M. A. and Lawrence, N. D. Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, (12):1425–1466, 2011.