Shared Stochastic Gaussian Process Latent Variable Models: A Multi-modal Generative model for Quasar spectra

Anonymous authors
Paper under double-blind review

Abstract

This work proposes a scalable probabilistic latent variable model based on Gaussian processes (Lawrence, 2004) in the context of multiple observation spaces. We focus on an application in astrophysics where it is typical for data sets to contain both observed spectral features as well as scientific properties of astrophysical objects such as galaxies or exoplanets. In our application, we study the spectra of very luminous galaxies known as quasars, and their properties, such as the mass of their central supermassive black hole, their accretion rate and their luminosity, and hence, there can be multiple observation spaces. A single data point is then characterised by different classes of observations which may have different likelihoods. Our proposed model extends the baseline stochastic variational Gaussian process latent variable model (GPLVM) (Lalchand et al., 2022) to this setting, proposing a seamless generative model where the quasar spectra and the scientific labels can be generated *simultaneously* when modelled with a shared latent space acting as input to different sets of Gaussian process decoders, one for each observation space. Further, this framework allows training in the missing data setting where a large number of dimensions per data point may be unknown or unobserved. We demonstrate high-fidelity reconstructions of the spectra and the scientific labels during test-time inference and briefly discuss the scientific interpretations of the results along with the significance of such a generative model.

1 Introduction

Many challenges in the contemporary physical sciences arise from the analysis of large scale, noisy, heteroscedastic and high-dimensional datasets (Clarke et al., 2016). Hence, there is increasing consensus that addressing the challenges posed by large scale data in the experimental pipelines for discovery, forecasting and prediction warrant scalable machine learning. Modern experiments in physics, chemistry and astronomy are capable of producing extremely complex high-dimensional data, but it is not just the sheer volume of the data but also the *velocity*, referring to the rate of production (Carleo et al., 2019) of data, which poses an additional challenge. Further, an additional axis is the variety or heterogeneity inherent in scientific datasets in the form of multiple outputs or observation spaces. For instance, images (pixels) can be accompanied with attributes like illumination, pose and resolution. In addition, some of these attributes or observed features are often unknown or unobserved, resulting in incomplete data sets with missing components. The multiple output setting is inadequately addressed in probabilistic machine learning literature; in this work we tackle precisely this setting proposing a multi-output scalable probabilistic model for a scientific application in astronomy.

A preponderance of literature in unsupervised learning focuses on the setting of a single large-scale homogeneous dataset, for instance, images, text, speech or continuous numerical values. Further, most parametric models also assume a complete dataset where each dimension of every data point is observed. Real-world data are often only partially observed and in some cases yield very sparse data matrices with majority of the dimensions missing (Corduneanu & Jaakkola, 2012). Robust unsupervised learning in these settings is a challenge and one needs to account for sensible epistemic uncertainties. In the running application we consider in this work, handling missing data during training is crucial as spectral pixels in real observations can be missing due to absorption features in the Earth's atmosphere or different wavelength coverage of the spectra when observed with different telescopes or spectrographs.

Dimensionality reduction is a standard precursor to further analysis in scientific datasets as the intrinsic dimensionality of the data might actually be quite low. It is usually possible to summarise the key axis of variation in the data with very few dimensions (Facco et al., 2017) side-stepping the curse of dimensionality and facilitating further downstream analysis. Projection techniques like PCA, multidimensional scaling (MDS) and independent component analysis (ICA) utilise eigenvalue decomposition (Murphy, 2012) while other non-linear techniques like t-SNE (van der Maaten & Hinton, 2008), SNE (van der Maaten, 2009) and UMAP (McInnes et al., 2020) construct a probability distribution on high dimensional points and replicate a similar distribution on low-dimensional points iteratively using the KL divergence. These methods however, are not generative in their traditional incarnation, they cannot be used to generate instances of the high-dimensional data points, hence, they are not very useful in many astrophysical settings where the ability to generate points in high-dimensional data space is crucial.

Generative latent variable models of late have supplanted traditional dimensionality reduction techniques as they offer the simultaneous benefits of a probabilistic interpretation and data generation while learning a faithful embedding of the high-dimensional training data in low-dimensional latent space. A generative probabilistic framework like the GPLVM (Lawrence, 2004) works by optimising the parameters of a Gaussian process *decoder* from a low dimensional latent space ($Z \in \mathbb{R}^{N \times Q}$) to high-dimensional data space ($X \in \mathbb{R}^{N \times D}$) such that $Q \ll D$ and points close in latent space are nearby in data space. Since the decoder is a non-parametric Gaussian process, the kernel function controls the inductive biases of the function mapping like smoothness and periodicity. There typically is no encoder mapping hence, these models are also called Gaussian process decoders. It is possible to additionally incorporate a back-constraint or an encoder which maps from the data to latent space putting GPLVMs on the same footing as variational auto-encoders (VAEs; Lawrence & Quiñonero Candela, 2006; Bui & Turner, 2015). This amortises the cost of variational inference in very large-scale datasets but we don't employ this setting as it is not straightforwardly applicable to missing data contexts.

This work proposes a novel formulation of the GPLVM based on the idea of a shared latent space. The earlier work by Ek (2009) was the first to propose the idea of a shared data generation process but precluded truly scalable inference due to the standard $\mathcal{O}(N^3)$ scaling. We extend this framework in two important ways. First, we show that the shared GPLVM is compatible with stochastic variational inference (SVI) (Hoffman et al., 2013) where we derive a joint evidence lower bound which factorises across multiple observation spaces due to conditional independence but share predictive strength though inducing locations and latent variables. Secondly, we train the entire model in presence of missing dimensions in one or both of the observation spaces.

We demonstrate this scalable model in an astrophysical application using data of quasars. Quasars are the most luminous galaxies in the universe, powered by accretion onto a central supermassive black hole (SMBH) with millions to billions of solar masses in size. Understanding the formation, growth, and evolution across cosmic time of quasars and their SMBHs is one of the major goals of observational cosmology today. To this end, precise measurements of the physical properties of quasars are crucial, but they typically demand very expensive and time-intensive observations, as multiple epochs are needed to accurately determine, for instance, the quasar's SMBH mass. The high-dimensional data used in this work contains $\sim 22,000$ quasars from the Sloan Digital Sky Survey (SDSS; Lyke et al., 2020; Wu & Shen, 2022) with high-quality spectral information (binned to 590 spectral dimensions/pixels) along with four scientific labels per quasar, i.e. their black hole mass, luminosity, redshift and so-called Eddington ratio – a measure of the quasar's accretion rate. By modelling the spectra and scientific labels through a generative model acting on a shared latent space we aim to reason about the physical properties of the quasars just through its "single-epoch" spectral information, thus circumventing the time-intensive multi-epoch observations. Earlier work on applying probabilistic generative modelling using a GPLVM to high-dimensional quasar spectra (Eilers et al., 2022) have been constrained on scalability and examine less than 50 astronomical objects. We demonstrate our framework on datasets $> 400\times$ bigger. We summarise our key contributions below:

Contributions We propose a probabilistic generative framework called the *Shared stochastic GPLVM* which is designed for use cases with multiple outputs/observation spaces. We seamlessly account for missing dimensions both at training and test time due to the probabilistic nature of the model. We demonstrate through astrophysical experiments that we can reconstruct previously unseen spectral pixels to a high degree of fidelity, interpolate missing or unobserved spectral regions and predict scientific labels. Crucially, we demonstrate that it is possible to share predictive strength by learning a common latent variable space Z across multiple-outputs (X, Y) where $Y \in \mathbb{R}^{N \times L}$ is an additional observation space with L dimensions. In this way we indirectly model the relationships and correlation structure between the different observation spaces. We demonstrate this concretely with an experiment where we generate/predict all the scientific attributes at test-time by using latent variables (Z) only informed by the quasar spectra (X), we can denote this cross modal prediction as $X \to Z \to Y$. To the best of our knowledge, predicting multiple outputs with stochastic variational GPs for scalability is methodologically novel. Further, this is the first demonstration of a scalable probabilistic latent variable model in astrophysical settings. In Section 8 we discuss how the development of these unsupervised learning frameworks can instigate novel insights into some of the major open questions in astronomy today.

2 Stochastic Variational GPLVM with a Shared latent space

In this section we first describe background on the stochastic variational GPLVM (Lalchand et al., 2022). We then develop the idea of a shared latent space and inducing points within the stochastic variational GPLVM framework. The fundamental contribution of this work is to develop an inference scheme to show that a shared latent space does not preclude scalable inference through SVI.

2.1 SV-GPLVM: Stochastic Variational GPLVM

In the traditional formulation underlying GPLVMs we have a training set comprising of N D-dimensional real valued observations $X \equiv \{\boldsymbol{x}_n\}_{n=1}^{N} \in \mathbb{R}^{N \times D}$. These data are associated with N Q-dimensional latent variables, $Z \equiv \{\boldsymbol{z}_n\}_{n=1}^{N} \in \mathbb{R}^{N \times Q}$ where $Q \ll D$ provides dimensionality reduction (Lawrence, 2004). The forward mapping $(Z \longrightarrow X)$ is governed by GPs independently defined across dimensions D. The sparse GP formulation describing the data is as follows:

$$
\begin{aligned}
p(Z) &= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{z}_n; \mathbf{0}, \mathbb{I}_Q), \\
p(F|U, Z, \theta) &= \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{f}_d; K_{nm} K_{mm}^{-1} \boldsymbol{u}_d, Q_{nn}), \\
p(X|F, Z) &= \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}(x_{n,d}; \boldsymbol{f}_d(\boldsymbol{z}_n), \sigma_x^2),
\end{aligned}
\tag{1}
$$

where $Q_{nn} = K_{nn} - K_{nm} K_{mm}^{-1} K_{mn}$, $F \equiv \{\boldsymbol{f}_d\}_{d=1}^{D}$, $U \equiv \{\boldsymbol{u}_d\}_{d=1}^{D}$ and $\boldsymbol{x}_d$ is the d^{th} column of X. K_{nn} is the covariance matrix corresponding to a user chosen positive-definite kernel function $k_\theta(x, x')$ evaluated on latent points $\{\boldsymbol{x}_n\}_{n=1}^{N}$ and parameterised by shared hyperparameters θ. The inducing variables per dimension $\{\boldsymbol{u}_d\}_{d=1}^{D}$ are distributed with a GP prior $\boldsymbol{u}_d|\tilde{Z} \sim \mathcal{N}(\boldsymbol{u}_d; \mathbf{0}, K_{mm})$ computed on inducing input locations $[\tilde{\boldsymbol{z}}_1, \dots \tilde{\boldsymbol{z}}_m]^T \equiv \tilde{Z} \in \mathbb{R}^{M \times Q}$ which live in latent space with Z and have dimensionality Q (matching $\boldsymbol{z}_n$).

The crux of SVI applied to sparse variational GPs as proposed in the seminal work of Hensman et al. (2013) is that we can variationally integrate out $\boldsymbol{u}_d$ by learning their variational distributions $q(\boldsymbol{u}_d) \sim \mathcal{N}(\boldsymbol{m}_d, S_d)$ numerically using stochastic gradient methods. Essentially, by keeping the representation of $\boldsymbol{u}_d$ uncollapsed. While (Hensman et al., 2013) proposed SVI for GP regression, (Lalchand et al., 2022) extended this work to GPLVMs where the inputs Z to the GPs are unobserved and each dimension of the high-dimensional output space $\boldsymbol{x}_d$ is modelled by an independent GP $\boldsymbol{f}_d$ with shared kernel hyperparameters.

Succinctly, posterior inference entails minimising the KL-divergence between the posterior over unknowns $p(F, U, Z|X)$ and the variational approximation $q(F, U, Z)$. When the variational approximation admits the factorisation below as in (Titsias, 2009):

$$q(F, U, Z) \approx p(F|U, Z)q(U)q(Z) = \prod_{d=1}^{D} p(\boldsymbol{f}_d|\boldsymbol{u}_d, Z)q(\boldsymbol{u}_d)\prod_{n=1}^{N} q(\boldsymbol{z}_n) \tag{2}$$

the evidence lower bound (ELBO) can be derived as,

$$\begin{aligned} \mathrm{KL}(q(F, U, Z)||p(F, U, Z|X)) &= \int p(F|U, Z)q(U)q(Z)\log\frac{p(F|U, Z)q(U)q(Z)}{p(F, U, Z|X)}dFdUdZ \\ &= -\underbrace{\int p(F|U, Z)q(U)q(Z)\log\frac{p(X|F, Z)\cancel{p(F|U, Z)}p(U)p(Z)}{\cancel{p(F|U, Z)}q(U)q(Z)}dFdUdZ}_{\text{ELBO}} + \log p(X) \end{aligned} \tag{3}$$

Minimising the KL is tantamount to maximising the ELBO as is clear from Eq. (3), this yiels the evidence lower bound,

$$\log p(X) \geq \mathbb{E}_{q(.)}[\log p(X|F, Z)] - \mathrm{KL}(q(Z)||p(Z)) - \mathrm{KL}(q(U)||p(U)) \tag{4}$$

For brevity we suppress the conditioning over the inducing inputs $\tilde{Z}$ in the prior $p(U)$ and the kernel hyperparameters θ in $p(F|U, Z)$.

If we choose to optimize the latent variables Z as point estimates rather than variationally integrate them out (basically we do not introduce $q(Z)$) we end up with the following simplification,

$$p(X) \geq \int p(F|U, Z)q(U)\log\frac{p(X|F, Z)p(U)p(Z)}{q(U)}dFdU = \mathcal{L}_x \tag{5}$$

Re-writing the lower bound as a sum of terms across data points N and outputs/dimensions D we get,

$$\mathcal{L}_x = \sum_{n,d}\langle\log p(x_{n,d}|\boldsymbol{f}_d, \boldsymbol{z}_n, \sigma_x^2)\rangle_{q(.)} - \sum_d \mathrm{KL}(q(\boldsymbol{u}_d)||p(\boldsymbol{u}_d|\tilde{Z})) + \sum_n \log p(\boldsymbol{z}_n) \tag{6}$$

Latent point estimates $\{\boldsymbol{z}_n\}_{n=1}^N$ can be learnt along with θ and variational parameters $(\tilde{Z}, \boldsymbol{m}_d, S_d)$ by taking gradients of the ELBO in Eq. (6). An important constraint however is that this formulation assumes a single kernel matrix (single set of kernel hyperparameters) underlying all the D independent GPs. In the next section, we introduce the idea of an additional observation space with L dimensions and how they can be modelled by their own stack of independent GPs $\boldsymbol{f}_l$ and learn their own set of hyperparameters for additional flexibility but share the latent embedding Z and inducing inputs $\tilde{Z}$ to model correlations between the different output spaces.

2.2 Shared joint variational lower bound

In the astrophysical application we focus on in this work we have two observation spaces corresponding to N quasars. We denote the quasar spectra (pixels) with the matrix $X \in \mathbb{R}^{N\times D}$ and the scientific labels corresponding to the N objects with $Y \in \mathbb{R}^{N\times L}$. The GPLVM construction models each column (pixel dimension and label dimension) with an independent GP, with the GPs corresponding to the pixel dimensions $\{f_d\}_{d=1}^D$ and label dimensions $\{f_l\}_{l=D+1}^L$ modelled with their own independent kernels and kernel hyperparameters, θ_x and θ_y. Within each observation space the kernel hyperparameters are shared, so we learn two sets of hyperparameters corresponding to two observation spaces.

$$f_d \sim \mathcal{GP}(0, k_{\theta_x}) \qquad f_l \sim \mathcal{GP}(0, k_{\theta_y}) \tag{7}$$

The priors over the function values are given by,

$$p(\boldsymbol{f}_d|\theta_x) = \mathcal{N}(\mathbf{0}, K_{nn}^{(d)}) \tag{8}$$

$$p(\boldsymbol{f}_l|\theta_y) = \mathcal{N}(\mathbf{0}, K_{nn}^{(l)}) \tag{9}$$

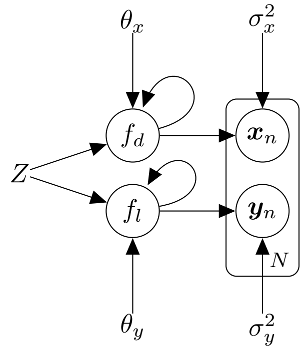

Figure 1: The graphical model of the shared GPLVM with two sets of independent GPs and their respective hyperparameter sets.

where $K_{nn}^{(d)}$ and $K_{nn}^{(l)}$ denote the $N \times N$ kernel matrices which rely on their own set of hyperparameters. The two observation spaces also yield two data likelihoods given by,

$$p(X|f_{1:D}, Z) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{f}_d, Z) = \prod_{n=1}^{N}\prod_{d=1}^{D} \mathcal{N}(x_{n,d}; f_d(\boldsymbol{z}_n), \sigma_x^2) \tag{10}$$

$$p(Y|f_{D+1:L}, Z) = \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{f}_l, Z) = \prod_{n=1}^{N}\prod_{l=D+1}^{L} \mathcal{N}(y_{n,l}; f_l(\boldsymbol{z}_n), \sigma_y^2) \tag{11}$$

In the absence of sparsity the log-marginal likelihood of the joint model compartmentalises nicely due to the assumed factorisation in the likelihoods. We marginalise out the latent function values $f_{1:D}$ and $f_{D+1:L}$ per dimension,

$$\begin{aligned} p(X, Y|\theta_x, \theta_y, Z) &= \int_{1:D} \dots \int_{D+1:L} p(X|f_{1:D}, Z)p(Y|f_{D+1:L}, Z)p(\boldsymbol{f}_d|\theta_x)p(\boldsymbol{f}_l|\theta_y)d\boldsymbol{f}_{1:D}d\boldsymbol{f}_{D+1:L} &(12)\\ &= \int_{1:D} p(X|f_{1:D})p(\boldsymbol{f}_d|\theta_x)d\boldsymbol{f}_{1:D}\int_{D+1:L} p(Y|f_{D+1:L})p(\boldsymbol{f}_l|\theta_y)d\boldsymbol{f}_{D+1:L} &(13)\\ &= \prod_{d=1}^{D} p(\boldsymbol{x}_d|\theta_x)\prod_{l=D+1}^{L} p(\boldsymbol{y}_l|\theta_y) = \prod_{d=1}^{D}\mathcal{N}(\boldsymbol{0}, K_{nn}^{(d)} + \sigma_x^2)\prod_{l=D+1}^{L}\mathcal{N}(\boldsymbol{0}, K_{nn}^{(l)} + \sigma_y^2) &(14)\end{aligned}$$

where $\boldsymbol{x}_d$ and $\boldsymbol{y}_l$ denote a single column/dimension of the observation spaces X and Y. The log marginal likelihood objective is then given by the following,

$$\log p(X, Y|\theta_x, \theta_y, Z) = \sum_{d=1}^{D} \log p(\boldsymbol{x}_d|\theta_x, Z) + \sum_{l=D+1}^{L} \log p(\boldsymbol{y}_l|\theta_y, Z) \tag{15}$$

In order to induce sparsity we introduce inducing variables $\boldsymbol{u}_d$ and $\boldsymbol{u}_l$ for each individual dimension in the observation spaces; however, they are underpinned by shared inducing inputs $\tilde{Z}$ which live in the shared latent space Z and share the same dimensionality, Q. With sparse GPs each of the terms in the decomposition above can be bounded by $\mathcal{L}_x$, while the inducing points $\tilde{Z}$ can be shared between the terms yielding the joint evidence lower bound.

$$\begin{aligned} \log p(X, Y|\theta_x, \theta_y, Z) \geq &\sum_{n,d}\langle \log p(x_{n,d}|\boldsymbol{f}_d, \boldsymbol{z}_n, \sigma_x^2)\rangle_{q(.)} - \sum_{d} \mathrm{KL}(q(\boldsymbol{u}_d)||p(\boldsymbol{u}_d)) \qquad (16)\\ &+ \sum_{n,l}\langle \log p(y_{n,l}|\boldsymbol{f}_l, \boldsymbol{z}_n, \sigma_y^2)\rangle_{q(.)} - \sum_{l} \mathrm{KL}(q(\boldsymbol{u}_l)||p(\boldsymbol{u}_l)) + \log p(Z)\end{aligned}$$

Overall we optimise the shared variaitonal lower bound w.r.t kernel hyperparameters for the two groups of GPs, θ_x and θ_y, variational parameters $\{\boldsymbol{m}_d, S_d\}_{d=1}^{D}$ and $\{\boldsymbol{m}_l, S_l\}_{l=D+1}^{L}$, and a single set of shared latent point estimates Z and inducing inputs $\tilde{Z}$. We include the full training algorithm in Algorithm 1.

3 Predictions & Reconstructions

High-dimensional points can arrive in different formats for the unseen data, either we observe both the modalities $\{\boldsymbol{x}^*, \boldsymbol{y}^*\}$ where $\boldsymbol{x}^* = [x_1^*, \ldots, x_d^*]^T$ and similarly for $\boldsymbol{y}^*$ or only one of the modalities with the other one missing i.e. $\{\boldsymbol{x}^*\}$ or $\{\boldsymbol{y}^*\}$ only. The prediction exercise then entails inferring the latent $\boldsymbol{z}^*$ corresponding to the unseen test point.

Since the GPLVM is a decoder only model, we can't obtain the latent embedding $\boldsymbol{z}^*$ deterministically, instead we re-optimise the ELBO with the additional test data point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ while keeping all the global and model hyperparameters frozen at their trained values. Note that since the ELBO factorises across data points, $\mathcal{L}(\{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N}, \boldsymbol{x}^*, \boldsymbol{y}^*) = \sum_{n=1}^{N+1} \sum_{h=1}^{D+L} \mathcal{L}_{n,h}$, the gradients to derive the new latent point $\boldsymbol{z}^*$ only depends on the respective component terms connected to the data point.

Once we infer $\boldsymbol{z}^*$ we can compute the full reconstruction distributions (we may be interested in this if the unseen data point $\boldsymbol{y}^*$ had any missing dimensions, for example $\boldsymbol{y}^* = [y_1^*, y_2^*, ?, ?, y_5^*, \ldots, y_l^*]^T$) which are just the GP posterior predictive for each column or dimension, without loss of generality, for dimension y_l^*:

$$\begin{aligned} p(y_l^*|\boldsymbol{z}^*) &= \int p(y_l^*|\boldsymbol{f}_l, \boldsymbol{u}_l, \boldsymbol{z}^*) p(\boldsymbol{f}_l, \boldsymbol{u}_l|\boldsymbol{y}_l) d\boldsymbol{f}_l d\boldsymbol{u}_l \\ &= \int p(y_l^*|\boldsymbol{f}_l, \boldsymbol{u}_l, \boldsymbol{z}^*) p(\boldsymbol{f}_l|\boldsymbol{u}_l) q(\boldsymbol{u}_l) d\boldsymbol{f}_l d\boldsymbol{u}_l \\ &= \int p(y_l^*|\boldsymbol{u}_l, \boldsymbol{z}^*) q(\boldsymbol{u}_l) d\boldsymbol{u}_l \end{aligned} \tag{17}$$

where $p(y_l^*|\boldsymbol{u}_l, \boldsymbol{z}^*) = \mathcal{N}(K_{*m}K_{mm}^{-1}\boldsymbol{u}_l, K_{**} - K_{*m}K_{mm}^{-1}K_{m*} + \sigma_y^2)$ and $q(\boldsymbol{u}_l) = \mathcal{N}(\boldsymbol{m}_l^*, S_l^*)$ refers to the optimised variational distribution. The final integral is tractable and gives the following form:

$$p(y_l^*|\boldsymbol{z}^*) = \mathcal{N}(K_{*m}K_{mm}^{-1}\boldsymbol{m}_l^*, K_{*m}K_{mm}^{-1}(S_l^* - K_{mm})K_{mm}^{-1}K_{m*} + \sigma_y^2) \tag{18}$$

and similarly for any x_d^*. For cross-modal reconstruction (where we only observe one modality of unseen data) the latent $\boldsymbol{z}^*$ acts as the information bottleneck, hence, the same posterior predictive distributions can be derived, $\boldsymbol{x}^* \rightarrow \boldsymbol{z}^* \longrightarrow p(y_l^*|\boldsymbol{z}^*) \forall l = D+1, ...L$.

4 Schematic of the Model

In Fig. 2 we present a schematic of the model architecture with two observation spaces (X, Y), the corresponding stacks of individual GPs $\{f_d\}$ and $\{f_l\}$ which model the individual columns of the spectra X and scientific attributes Y, respectively, and the low-dimensional latent space Z. The dimensionality of the latent and observation spaces are denoted by Q, D, L, respectively, and N denotes the number of objects / data points (quasars). Note that the correlation between the two observation spaces are not explicitly but implicitly modelled through a shared latent space. Generating a single data point $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ (a row across X and Y) entails a forward pass through the GPs, where $\boldsymbol{x}_n = [\ldots, x_{nd}, \ldots]$ is generated as $[f_1(\boldsymbol{z}_n), f_2(\boldsymbol{z}_n), \ldots, f_D(\boldsymbol{z}_n)]$ and $\boldsymbol{y}_n = [\ldots, y_{nl}, \ldots]$ is generated as $[f_{D+1}(\boldsymbol{z}_n), f_{D+2}(\boldsymbol{z}_n), \ldots, f_{D+L}(\boldsymbol{z}_n)]$.

5 Algorithm

We enclose the pseudo-code in Algorithm 1 for stochastic variational inference in the context of the shared model for clarity. Let $\mathcal{L}_x$ and $\mathcal{L}_y$ denote the ELBO's for each of the observation spaces and let $\mathcal{L}_x^{(B)}$ and $\mathcal{L}_y^{(B)}$ denote the ELBOs formed with a randomly drawn mini-batch of the data (across all dimensions). For a

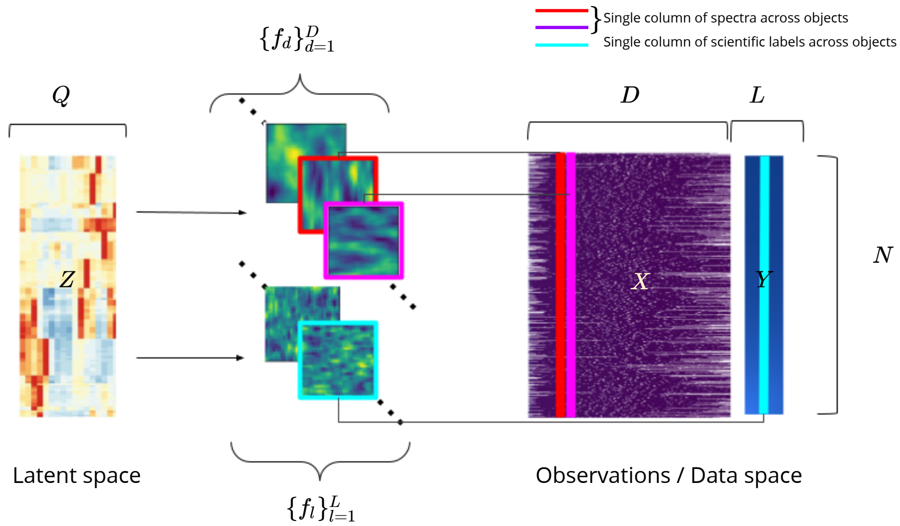

Figure 2: Shared GPLVM with multiple observation spaces. The blocks on the right-hand side denote the double observation spaces (X, Y) of quasar spectra and scientific labels respectively. In the center are two stacks of GPs, one for each observation space which control the data generation process through the shared latent space. In the figure above we assume $Q = 2$ (for ease of visualisation) since we denote the GPs are two dimensional surfaces, however, typically Q can be higher than 2 corresponding to higher dimensional GPs.

mini-batch (subset) of the data $X_B \subset X$, the mini-batch ELBO is given by,

$$\mathcal{L}_x \simeq \mathcal{L}_x^{(B)} = \frac{N}{B}\left(\sum_{b,d}\langle \log p(x_{b,d}|\boldsymbol{f}_d, z_b, \sigma_x^2)\rangle_{q(.)} + \sum_b \log p(\boldsymbol{z}_b)\right) - \sum_d \mathrm{KL}(q(\boldsymbol{u}_d)||p(\boldsymbol{u}_d|\tilde{Z})) \tag{19}$$

where the scaling term is important for the mini-batch ELBO to be an estimator of the full-dataset ELBO.

5.1 Computational Cost

The training cost of the canonical stochastic variational GPLVM is dominated by the number of inducing points $\mathcal{O}(M^3 D)$ (free of N) where $M \ll N$ and D is the data-dimensionality (we have D GP mappings f_d, one per output dimension). The practical algorithm is made further scalable with the use of mini-batched learning. In our shared model with two sets of GPs the dynamics of the training cost are the same except that they go up linearly in the number of additional dimensions (L), making the cost $\mathcal{O}(M^3(D+L))$. The number of global variational parameters to be updated in each step (parameters of $q(U)$) is $MQ + M(D+L) + M^2(D+L)$, where MQ are the M Q-dimensional inducing inputs $\tilde{Z}$ (shared), $M(D+L)$ is the size of the mean parameters of the inducing variables $\boldsymbol{u}_d$, $\boldsymbol{u}_l$ and $M^2(D+L)$ are the full-rank covariances of the inducing variables. The local variational parameters Z (the latent embedding shared across GPs) are of size NQ and model hyperparameters (kernel hyperparameters) are of size $2Q+4$, which account for Q input lengthscales, a scalar signal variance and noise variance per GP group $\{f_d\}$ and $\{f_l\}$. We use the squared exponential kernel with automatic relevance determination across both sets of GPs.

6 Related methodological work

In this section we present related work in multi-output Gaussian processes in more detail. Canonical multi-output Gaussian processes almost always refer to the supervised framework or multi-output regression (Álvarez et al., 2010) where the targets $\{\boldsymbol{x}_d(\boldsymbol{z})\}_{d=1}^{D}$ are multi-dimensional, continuous and Gaussian distributed corresponding to inputs $\boldsymbol{z} \in \mathbb{R}^p$. The main focus is on defining a suitable cross-covariance function between the outputs, this allows treatment of the outputs as a single GP with a suitable covariance function (Alvarez et al., 2012). The intrinsic and linear coregionalisation models (Goovaerts, 1993; Journel & Huijbregts, 1976) are a popular approach where each output is modelled as a weighted sum of shared latent functions (GPs)

Algorithm 1: Shared Stochastic GPLVM for Quasar Spectra

TRAINING FRAMEWORK

Input: ELBO objective $\mathcal{L} = \mathcal{L}_x + \mathcal{L}_y$, gradient based optimiser `optim()`, observation spaces X (spectra) and Y (scientific labels)
Initial model params:
$\theta = (\theta_x, \theta_y)$ (covariance hyperparameters for GP mappings $f_{1:D}, f_{D+1:L}$),
$\sigma^2 = (\sigma_x^2, \sigma_y^2)$ (variance of the noise model for each likelihood),
$Z \equiv \{\boldsymbol{z}_n\}_{n=1}^N$ (point estimates for latent embedding)
Initial variational params:
$\tilde{Z} \in \mathbb{R}^{M \times Q}$ (inducing locations),
$\lambda = \{m_h, S_h\}_{h=1}^{D+L}$ (global variational params for inducing variables per dimension $\boldsymbol{u}_h$),

while *not converged* **do**

- Choose a random mini-batch of the data from both the observation spaces $X_B \subset X, Y_B \subset Y$.
- Form a mini-batch estimate of the ELBO: $\mathcal{L}_x^{(B)} + \mathcal{L}_y^{(B)}$
- Gradient step for global parameters $\boldsymbol{g} \leftarrow \nabla_{\theta, \sigma^2, \tilde{Z}, \lambda}(\mathcal{L}_x^{(B)} + \mathcal{L}_y^{(B)})$
- Gradient step for local parameters $\boldsymbol{l} \leftarrow \nabla_{Z_B}(\mathcal{L}_x^{(B)} + \mathcal{L}_y^{(B)})$ (where Z_B are the latent embeddings corresponding to points in the mini-batch)
- Update all parameters $\tilde{Z}, \theta, \sigma^2, \lambda, Z_B \equiv \{\boldsymbol{z}_b\}_{n=1}^B \longleftarrow$ `optim()` using gradients $\boldsymbol{g}, \boldsymbol{l}$

end
return $\theta, \sigma^2, \tilde{Z}, \lambda, Z$

PREDICTION FRAMEWORK
(Predict Z^* corresponding to unseen X^*, Y^*)

Input: Trained global and local parameters $\theta, \sigma^2, \tilde{Z}, \lambda, Z$, unseen observation spaces X^* (spectra) and Y^* (scientific labels).
1. Initialise latent embedding $Z^* \equiv \{\boldsymbol{z}_{n^*}\}_{n^*=1}^{N^*}$ corresponding to unseen points.
2. Extend the joint ELBO to include terms corresponding to the N^* additional data points.
$\mathcal{L}_x^* \leftarrow \mathcal{L}_x + \sum_{n^*,d} \langle \log p(x_{n^*,d} | \boldsymbol{f}_d, z_{n^*}, \sigma_x^2) \rangle_{q(.)} + \sum_{n^*} \log p(\boldsymbol{z}_{n^*})$
$\mathcal{L}_y^* \leftarrow \mathcal{L}_y + \sum_{n^*,l} \langle \log p(y_{n^*,l} | \boldsymbol{f}_l, z_{n^*}, \sigma_y^2) \rangle_{q(.)} + \sum_{n^*} \log p(\boldsymbol{z}_{n^*})$
$\mathcal{L}^* = \mathcal{L}_x^* + \mathcal{L}_y^*$
3. Freeze all global and local parameters except for Z^*
while *not converged* **do**

- Gradient step for Z^*: $\boldsymbol{l}^* \leftarrow \nabla_{Z^*} \mathcal{L}^*$
- Update $Z^* \leftarrow$ `optim()` using gradients $\boldsymbol{l}^*$.

end
return Z^*
(Note that the gradients of $\mathcal{L}_x$ and $\mathcal{L}_y$ with respect to Z^* are 0 and the only terms that are optimised are the additional terms corresponding to the new data points.)

where typically the number of latent processes is smaller than the number of outputs enabling efficiencies. To some extent multi-task learning with GPs can be viewed as an instance of multi-output learning where we want to avoid tabula rasa learning for each task and evolve a framework for sharing information between multiple tasks (Bonilla et al., 2007). Further, there is the simplistic paradigm where each output is modelled with its own independent single-output GP and no correlation between the outputs is assumed. This approach while easy to implement, is severely limited in its ability to jointly model the outputs.

In the unsupervised paradigm, the starting point is a high-dimensional data matrix $N \times D$. The conventional Gaussian process latent variable model (GPLVM) operates like a multi-output model by default where each column of the data is modelled by an independent GP on the same shared set of inputs, kernel function and hyperparameters. The GPLVM is a decoder only model and some of their prominent variants are the back-constrained GPLVM (Lawrence & Quiñonero Candela, 2006), supervised GPLVM (Jiang et al., 2012), discriminative GPLVM (Urtasun & Darrell, 2007) and the shared GPLVM (Ek, 2009). The latter considers the task of dealing with multiple views or observation spaces (each of which is high-dimensional). This work is built on the idea of a shared latent space underlying the multiple observation spaces similar to (Ek et al., 2007) but adapts it for scalable inference using stochastic variational inference (SVI). Stochastic variational

Metrics (⟶)	RMSE			
Models (⟶)	Baseline		Shared (ours)	
Attributes (↓)	1k	22k	1k	22k
Spectra	0.0731 ± 0.0020	0.1217 ± 5e-4	0.0707 ± 0.0033	**0.1210 ± 4e-4**
Blackhole Mass	0.1882 ± 0.0046	0.2846 ± 0.0037	**0.1765 ± 0.0054**	**0.2453 ± 0.0014**
Bolometric Luminosity	0.1731 ± 0.0043	0.2562 ± 0.0034	0.1658 ± 0.0084	**0.2267 ± 0.0022**
Eddington Ratio	0.1707 ± 0.0024	0.2799 ± 0.0061	**0.1653 ± 0.0026**	**0.2336 ± 0.0021**

Table 1: Summary of test-time reconstruction abilities. Mean absolute error on denormalised data (± standard error of mean) evaluated on average of 5 splits with 75% of the data used for training in the 22k dataset and 90% of the data used for training in the 1k dataset. The shared model outperforms the baseline model in all the reconstruction tasks for the larger dataset. In the smaller dataset the performance improvement is not statistically significant for the spectra reconstruction and the bolometric luminosity prediction.

GPLVM was proposed in (Lalchand et al., 2022) but only considered a single observation space as in a canonical GPLVM.

7 Experiments

In this section we demonstrate a range of experiments aimed at assessing the reconstruction quality of unseen quasar spectra and scientific attributes, as well as the robustness of uncertainty quantification by computing the negative log predictive density of unseen labels $-\log p(Y^*|Z^*)$ under the predictive distribution where Z^* has been informed by both modalities (spectra and labels) as well as just the spectra.

The data used in this work are quasar spectra observed as part of the Sloan Digital Sky Survey (SDSS) DR16 (Lyke et al., 2020). We chose all quasars with spectra that have a signal-to-noise ratio (SNR) per pixel > 10, which results in a total of 22844 quasar spectra. The observed spectra are shifted into the rest-frame wavelength space, re-binned onto a common wavelength grid, and flux normalized to unity at around 2500 Å. We mask strong absorption lines that might arise in the spectra due to foreground galaxies along our line-of-sight to the quasar, and are thus not intrinsic spectral features of the quasar. The four scientific labels for these quasars are (1) their SMBH mass, (2) their bolometric luminosity, i.e. the total power output across all electromagnetic wavelengths, (3) their redshift which denotes the factor by which the emitted wavelengths have been "stretched" due to the expansion of the universe, and (4) their Eddington ratio, which is a measure of the accretion and growth rate of the SMBH. All measurements were previously uniformly determined by Wu & Shen (2022). We conduct experiments across two data sets with 1k and 22k points (as an ablation to assess the performance with a smaller dataset). The 1k dataset was also derived from SDSS with the same SNR threshold. The "Baseline" model in the experiments refers to the canonical stochastic variational GPLVM (Lalchand et al., 2022) which treats multiple observation spaces using the same set of independent GPs learning a single set of kernel hyperparameters

7.1 Reconstructing Quasar spectra

We assess the quality of our probabilistic generative model in reconstructing unseen quasar spectra. At test-time we deal with spectra and scientific labels from unseen quasars denoted by $(X^*_{\text{gt}}, Y^*_{\text{gt}})$ (we use the index 'gt' to denote ground truth). The 2-step prediction learns the low-dimensional shared latent variables Z^* (point estimate per data point) which acts as input to the GP decoder predicting the corresponding spectra $Z^* \rightarrow X^*_{\text{est}}$. Note that the ground truth spectra contains several missing pixels (dimensions) and the probabilistic decoder provides a reasonable reconstruction at those locations. In Fig. 3 we visualise the reconstruction (posterior predictive mean) of four test quasars along with ground-truth measurements and 2σ uncertainty intervals. We achieve a remarkably good reconstruction estimates from the latents; further, the prediction intervals capture the ground spectra providing robust coverage at peaks and extrapolated regions.

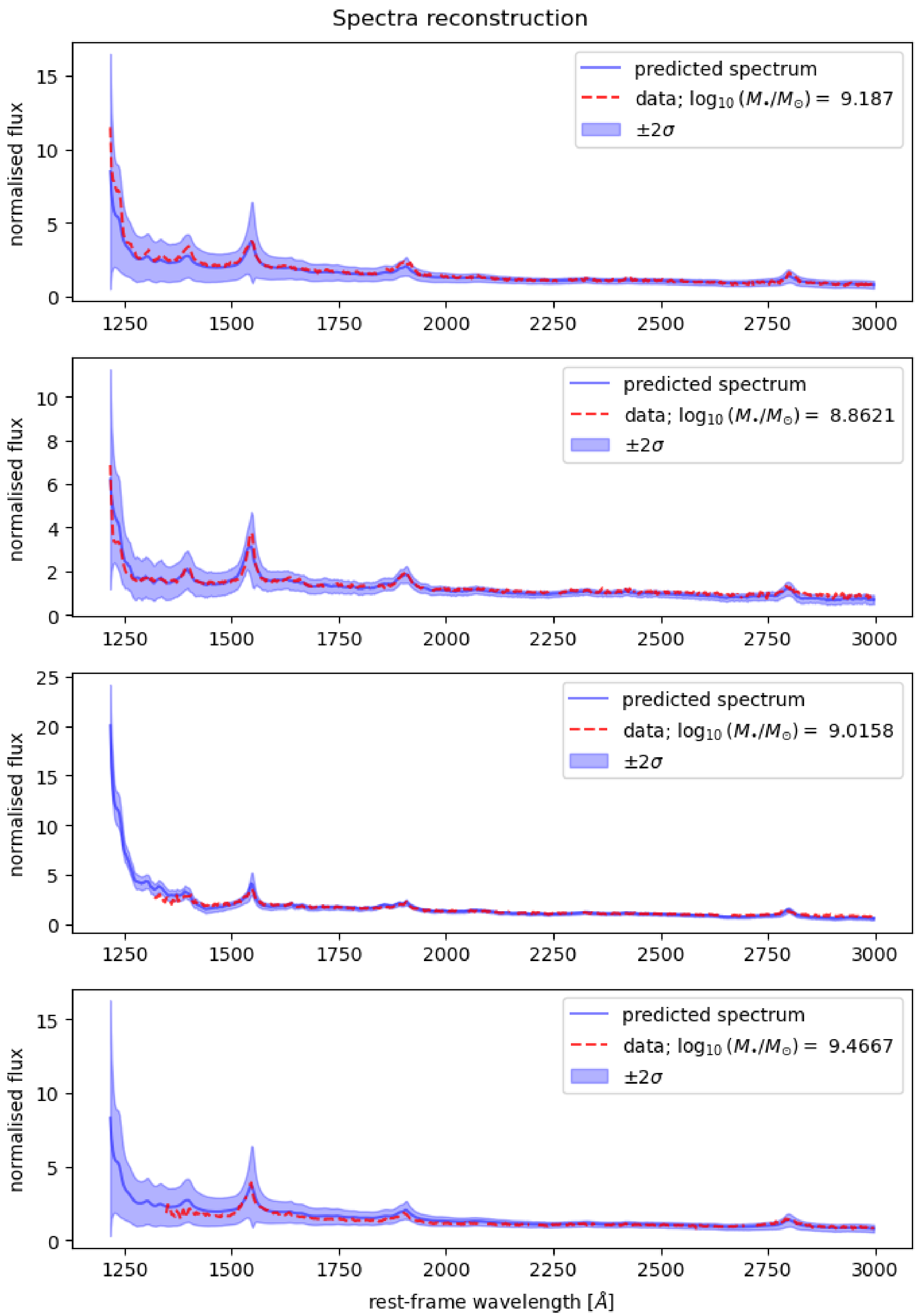

Figure 3: Reconstruction plots of unseen quasar spectra with $\pm 2\sigma$ predictions intervals. The blue curve denotes the posterior predictive mean at each dimension.

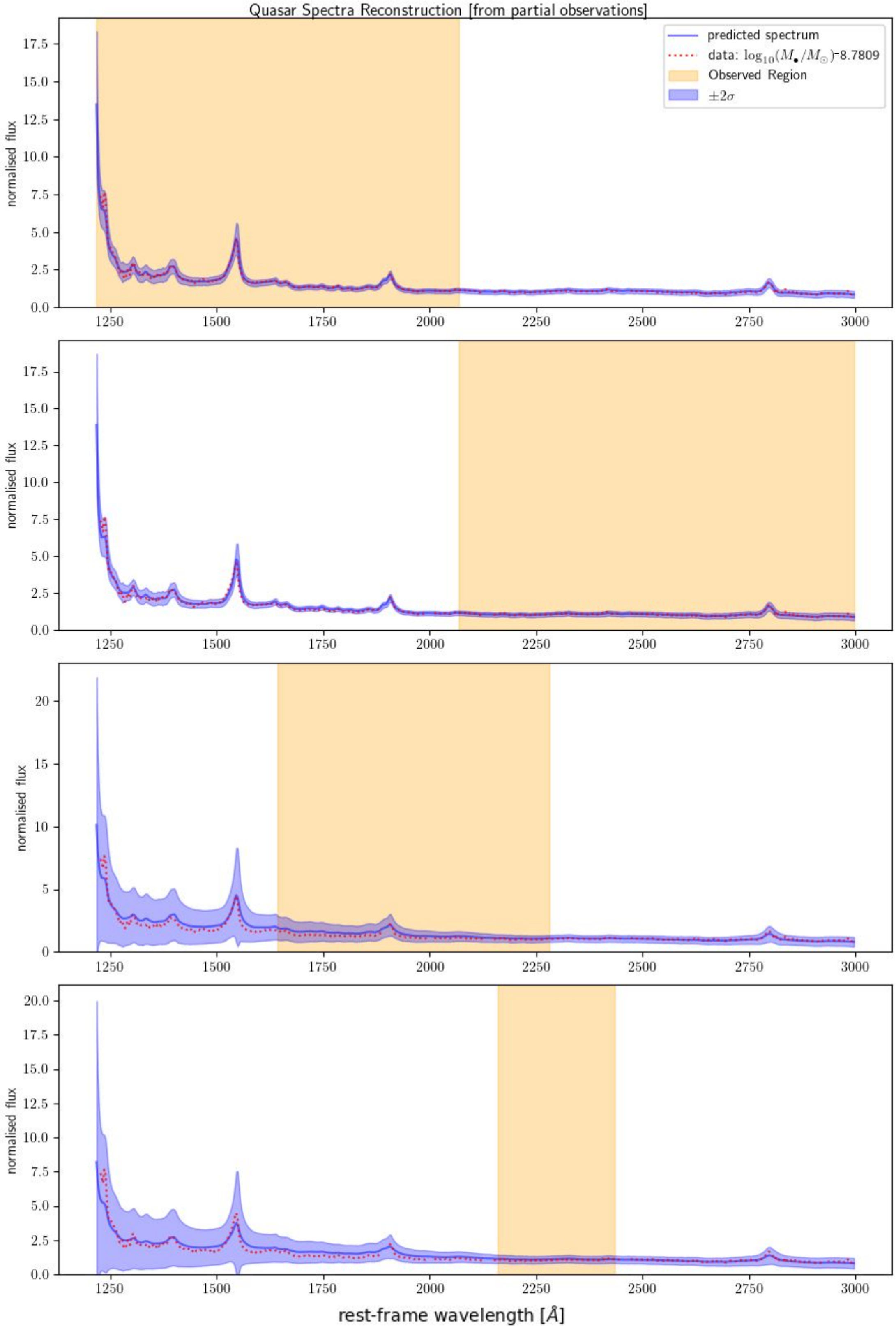

Figure 4: Reconstruction of a single spectra from the latent informed by a partially observed spectrum. The shaded orange regions denote the "observed" wavelength regions for this experiment. Note in the 3rd and 4th panels the 2σ prediction intervals are wider at the initial wavelengths as they were observed over a shorter and less informative wavelength window.

7.1.1 Reconstructing missing spectra

In this experiment we test the generative model's ability to learn from massively missing chunks of the spectra at test-time. We observe a partial window of the spectra in each plot (given by the shaded region in Fig. 4), hence the latent variables corresponding to these points are only informed by the observed region. We then reconstruct the whole spectra from the latent variables informed by the partial spectra. We enclose our results in Fig. 4. The reconstruction entails the inference steps: $X^*_{\text{partial}} \to Z^* \to X^*_{\text{full}}$. We note that the quality of the mean prediction deteriorates compared to the fully observed test point predictions. However, the coverage of the prediction intervals is robust even as we move away from the shaded observed regions. Uncertainty intervals are much higher at the unobserved regions, indicating highly sensible predictive behaviour. Furthermore, if the model is given a spectral region with very little information (e.g. in the bottom panel of Fig. 4, the shaded region contains only information about the quasar's continuum, but no emission lines are captured), the uncertainties increase significantly, as one would expect.

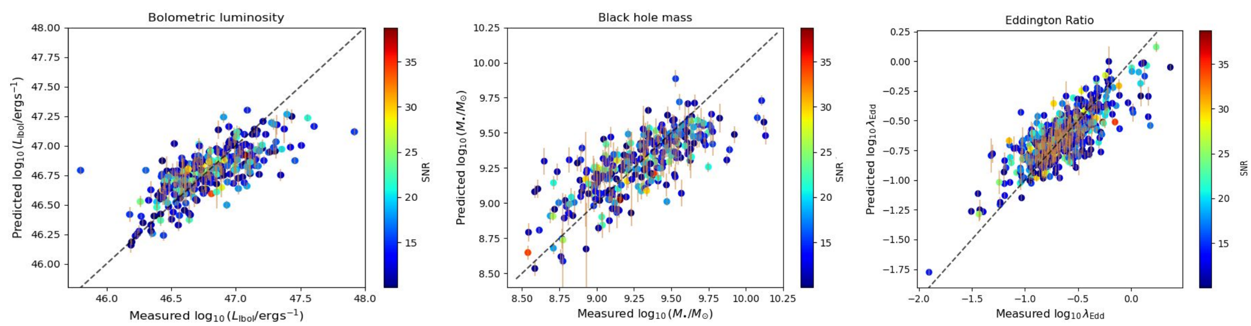

Figure 5: Scientific label prediction for the quasars' bolometric luminosity (left), black hole mass (middle) and Eddington ratio (right) colored by the SNR of their spectra based on unseen X^* **only**. The dashed black line (- - -) denotes the 1-to-1 line to aid visualisation of reconstruction accuracy. The vertical and horizontal errorbars (——) denotes posterior predictive standard deviation.

7.2 Predicting scientific labels *only* from spectra X^*

The L dimensions corresponding to the scientific labels in the dataset governed by their own GP decoders $\{f_l\}_{l=1}^{L}$ are a critical prediction quantity. The ability to reconstruct these quantities from learnt latent variables is an important test of the generalisation abilities of our model. Very often astronomers want to reason about the scientific attributes of quasars just by analysing their spectra. In this experiment we demonstrate precisely this use case where the latent variables Z^* are informed only by the spectra X^*, computing the cross-modal prediction entails learning Z^* from the ground-truth spectra and then using just Z^* to predict the scientific labels Y^*, succinctly, we can write these steps as: $X^* \to Z^* \to Y^*$.

In Fig. 5 we demonstrate the accuracy of our reconstructions by plotting each of the dimensions against ground truth held-out data. We show reconstructions for 200 test points sampled randomly from the full test set. Each point on the scatter denotes a quasar and the x-axis denotes the ground-truth measurement. The orange vertical error bars denote 2σ intervals computed by extracting the diagonals from the GP posterior predictive for each dimension. We can observe a high-degree of prediction accuracy across the three scientific labels and further, the reconstruction quality is robust and independent of the spectral signal-to-noise ratio (SNR) as there is no strong pattern of correlation between prediction quality and SNR. Note that we omit the reconstruction of the redshift label z, as we do not expect any redshift evolution in the spectral shape of quasars, but rather the GPLVM is likely learning where spectral pixels are measured vs. missing. Since we are using data from SDSS, all spectra have the same observed wavelength coverage, which is then shifted to the rest-frame wavelengths by dividing by $(1 + z)$, and thus the observed pixel dimensions in our data set contains the redshift information, rather than any spectral differences.

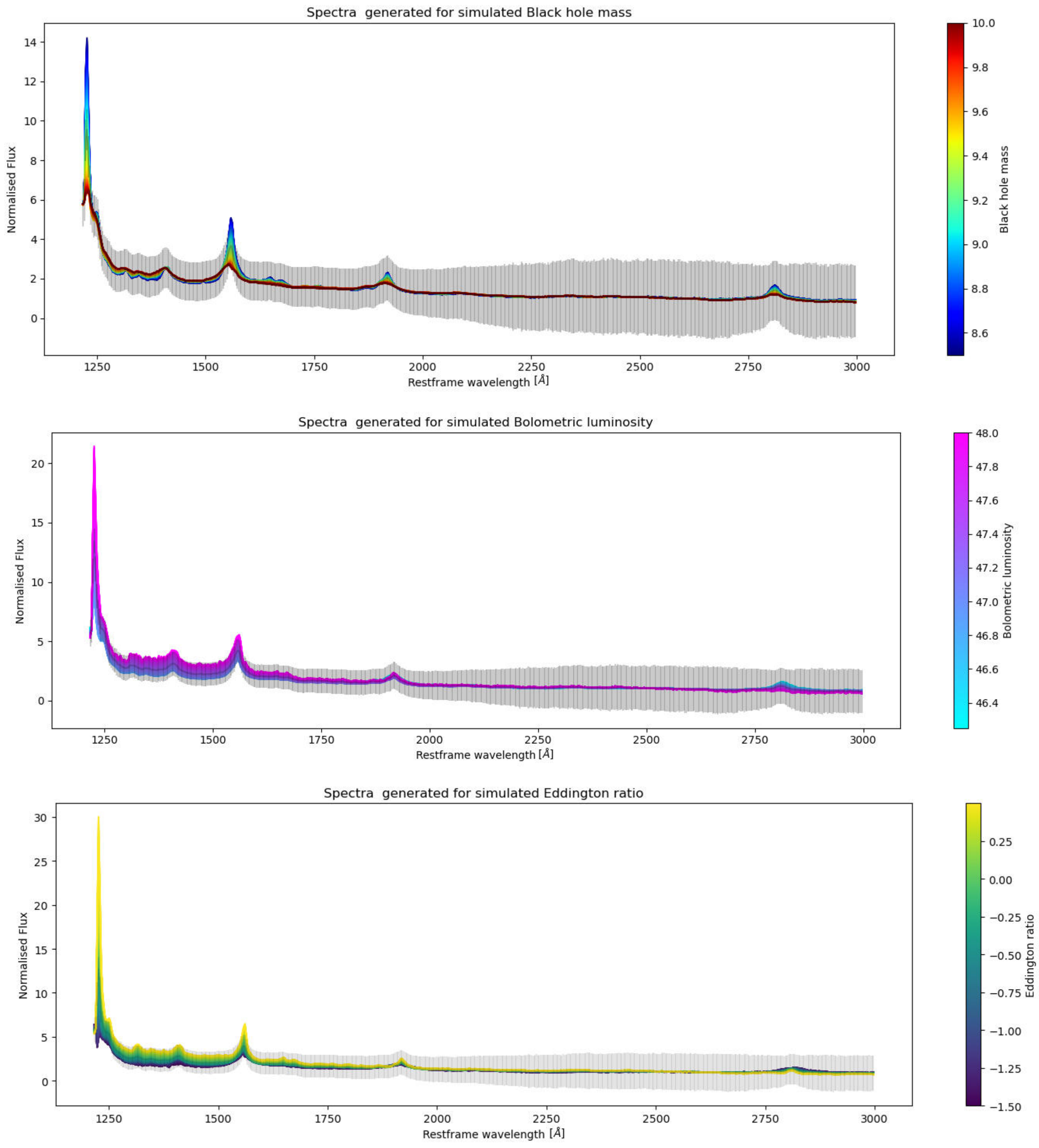

Figure 6: Generated quasar spectra for simulated labels for black hole mass, bolometric luminosity and Eddington ratio.

Experiment	Observation → Latent → Prediction	Black hole mass	Luminosity	Eddington Ratio
Fully observed	$(X^*_{\text{gt}}, Y^*_{\text{gt}}) \to Z^* \to Y^*_{\text{est}}$	0.4428	0.3823	0.3085
Spectra observed	$(X^*_{\text{gt}} \to Z^* \to Y^*_{\text{est}})$	0.4943	0.4339	0.3336

Table 2: Summary of test-time uncertainty quantification under the full and partial reconstruction framework. Median negative log predictive density (lower is better) on de-normalised data across 5 train/test splits. A higher NLPD indicates lower confidence in the predictions and the model indicates this when reconstructing the labels of a quasar just on the basis of its spectrum.

7.3 Generating spectra corresponding to synthesised labels: an ablation study

In this experiment we demonstrate the models ability to generate spectra corresponding to synthesized scientific labels. Concretely, we simulate artificial labels by systematically varying only one of the labels within a reasonable range in each plot in Fig. 6. The range of variation for each label is summarised by the colorbar in each plot, for instance in the black hole mass simulation we generate 100 spectra (X^*) corresponding to simulated labels (Y^*) of black hole masses in the range $\log_{10}(M_\bullet/M_\odot) = 7.9 - 10.0$; in order to ablate the influence of other labels (redshift, bolometric luminosity and the Eddington ratio) on the spectra we keep their values fixed to mean values computed from the training dataset. Succinctly, we can summarise the inference steps as the inverse of the one from the previous section: $Y^* \to Z^* \to X^*$. The ability for cross-modal prediction and generation is an important strength of our design.

Reassuringly, the model exhibits the expected behavior. For instance, the emission lines are broader for quasars with higher black hole masses, and the spectral dependency with bolometric luminosity shows e.g. in the Mg II line at ≈ 2800Å the well known Baldwin effect (Baldwin, 1977), which indicates that quasar spectra show a decreasing equivalent width of their UV and optical emission lines with increasing bolometric luminosity.

Further, the grey bars denote 2σ prediction intervals averaged across the 100 spectra at each dimension. The uncertainty intervals are wider at higher wavelengths as there is a high concentration of missing pixels in the training data at those wavelengths, hence, there is greater uncertainty about the generated spectra.

8 Scientific Interpretation and Significance

Our new generative model allows us to *simultaneously* model the spectral properties of quasars as well as their scientific labels, thus opening up novel possibilities to study the evolution of quasars across cosmic time and the formation and growth of SMBHs. In the following we will highlight just two possible and exciting future applications of this work.

Astronomers observe SMBHs with billions of solar masses in size in the center of very distant, high-redshift quasars, at a time when the universe is still in its infancy and only a few hundred Myr old. This rapid growth of SMBHs in the very short amounts of available cosmic time has been an open puzzle for decades, and it has been argued that very high accretion rates in excess of the theoretical upper limit, the so-called Eddington limit, are required to explain the rapid black hole growth. However, obtaining precise black hole mass measurements of quasars is challenging and time-intensive as it requires multi-epoch observations of certain emission lines in the quasar spectra (Peterson, 1993; Barth et al., 2015). This procedure becomes increasingly challenging for very distant, high-redshift quasars, as relativistic time-dilation effect require longer timespans of these observations. Furthermore, the traditionally used rest-frame optical emission lines to calibrate the black hole masses are unobservable with ground-based observatories, as these optical wavelengths have been shifted to the infrared at larger distances, which require space-based telescopes, such as NASA's newly launched James Webb Space Telescope. Our new model allows us to determine the masses of SMBHs for quasars at all redshifts from single-epoch data and does not require coverage of specific emission lines, since it can handle missing data. This opens up the possibilities to determine precise black hole masses and accretion rates for quasars across cosmic time to study the growth phases of SMBHs.

Additionally, we have shown that our model is able to predict also other physical properties of quasars such as their bolometric luminosities (see Fig. 5). This suggests that we can obtain a measurement of the quasars' absolute luminosities from their spectra alone, which enables us to use quasars as so-called "standard candles". Standard candles are incredibly valuable for astronomy, as knowing the luminosity of an object allows one to determine its distance. Famously, supernovae have been used as standard candles, which led to the Nobel Prize winning discovery of the expansion of our universe and the existence of dark energy (Riess et al., 1998). Our new model allows us to now also use quasars as standard candles, which are much more numerous than supernovae and can be probed to larger distances due to their on average much higher luminosities, enabling new constraints on the dark energy content and the Hubble constant of our universe in the future.

References

Mauricio Álvarez, David Luengo, Michalis Titsias, and Neil D Lawrence. Efficient multioutput gaussian processes through variational inducing kernels. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 25–32. JMLR Workshop and Conference Proceedings, 2010.

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Jack A. Baldwin. Luminosity Indicators in the Spectra of Quasi-Stellar Objects. *The Astrophysical Journal*, 214:679–684, June 1977. doi: 10.1086/155294.

Aaron J. Barth, Vardha N. Bennert, Gabriela Canalizo, Alexei V. Filippenko, Elinor L. Gates, Jenny E. Greene, Weidong Li, Matthew A. Malkan, Anna Pancoast, David J. Sand, Daniel Stern, Tommaso Treu, Jong-Hak Woo, Roberto J. Assef, Hyun-Jin Bae, Brendon J. Brewer, S. Bradley Cenko, Kelsey I. Clubb, Michael C. Cooper, Aleksandar M. Diamond-Stanic, Kyle D. Hiner, Sebastian F. Hönig, Eric Hsiao, Michael T. Kandrashoff, Mariana S. Lazarova, A. M. Nierenberg, Jacob Rex, Jeffrey M. Silverman, Erik J. Tollerud, and Jonelle L. Walsh. The Lick AGN Monitoring Project 2011: Spectroscopic Campaign and Emission-line Light Curves. *The Astrophysical Journal Supplement Series*, 217(2):26, April 2015. doi: 10.1088/0067-0049/217/2/26.

Edwin V Bonilla, Kian Chai, and Christopher Williams. Multi-task gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

Thang D. Bui and Richard E. Turner. Stochastic variational inference for Gaussian process latent variable models using back constraints. In *Black Box Learning and Inference NIPS workshop*, 2015.

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

P. Clarke, P. V. Coveney, A. F. Heavens, J. Jäykkä, B. Joachimi, A. Karastergiou, N. Konstantinidis, A. Korn, R. G. Mann, J. D. McEwen, S. de Ridder, S. Roberts, T. Scanlon, E. P. S. Shellard, and J. A. Yates. Big data in the physical sciences: challenges and opportunities. *ATI Scoping Report*, 2016.

Adrian Corduneanu and Tommi S Jaakkola. Continuation methods for mixing heterogenous sources. *arXiv preprint arXiv:1301.0562*, 2012.

Anna-Christina Eilers, David W Hogg, Bernhard Schölkopf, Daniel Foreman-Mackey, Frederick B Davies, and Jan-Torge Schindler. A generative model for quasar spectra. *The Astrophysical Journal*, 938(1):17, 2022.

Carl Henrik Ek. *Shared Gaussian process latent variable models*. PhD thesis, Citeseer, 2009.

Carl Henrik Ek, Philip H. S. Torr, and Neil D. Lawrence. Gaussian process latent variable models for human pose estimation. In *International workshop on machine learning for multimodal interaction*, pp. 132–143. Springer, 2007.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

P Goovaerts. Spatial orthogonality of the principal components computed from coregionalized variables. *Mathematical Geology*, 25:281–302, 1993.

James Hensman, Nicoló Fusi, and Neil D. Lawrence. Gaussian processes for big data. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI2013)*, 2013.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013. URL `http://jmlr.org/papers/v14/hoffman13a.html`.

Xinwei Jiang, Junbin Gao, Tianjiang Wang, and Lihong Zheng. Supervised latent linear gaussian process latent variable model for dimensionality reduction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(6):1620–1632, 2012.

Andre G Journel and Charles J Huijbregts. *Mining geostatistics*. 1976.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2013.

Vidhi Lalchand, Aditya Ravuri, and Neil D. Lawrence. Generalised GPLVM with Stochastic Variational Inference. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7841–7864. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/lalchand22a.html`.

Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*, pp. 329–336, 2004.

Neil D. Lawrence and Joaquin Quiñonero Candela. Local distance preservation in the GPLVM through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, pp. 513–520, 2006.

Brad W. Lyke, Alexandra N. Higley, J. N. McLane, Danielle P. Schurhammer, Adam D. Myers, Ashley J. Ross, Kyle Dawson, Solène Chabanier, Paul Martini, Nicolás G. Busca, Hélion du Mas des Bourboux, Mara Salvato, Alina Streblyanska, Pauline Zarrouk, Etienne Burtin, Scott F. Anderson, Julian Bautista, Dmitry Bizyaev, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Johan Comparat, Paul Green, Axel de la Macorra, Andrea Muñoz Gutiérrez, Jiamin Hou, Jeffrey A. Newman, Nathalie Palanque-Delabrouille, Isabelle Pâris, Will J. Percival, Patrick Petitjean, James Rich, Graziano Rossi, Donald P. Schneider, Alexander Smith, M. Vivek, and Benjamin Alan Weaver. The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. *The Astrophysical Journal Supplement Series*, 250(1):8, September 2020. doi: 10.3847/1538-4365/aba623.

Chao Ma, Wenbo Gong, José Miguel Hernández-Lobato, Noam Koenigstein, Sebastian Nowozin, and Cheng Zhang. Partial vae for hybrid recommender system. In *NIPS Workshop on Bayesian Deep Learning*, volume 2018, 2018a.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018b.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Bradley M. Peterson. Reverberation Mapping of Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, 105:247, March 1993. doi: 10.1086/133140.

Adam G. Riess, Alexei V. Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M. Garnavich, Ron L. Gilliland, Craig J. Hogan, Saurabh Jha, Robert P. Kirshner, B. Leibundgut, M. M. Phillips, David Reiss, Brian P. Schmidt, Robert A. Schommer, R. Chris Smith, J. Spyromilio, Christopher Stubbs, Nicholas B. Suntzeff, and John Tonry. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *The astronomical journal*, 116(3):1009–1038, September 1998. doi: 10.1086/300499.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Raquel Urtasun and Trevor Darrell. Discriminative Gaussian process latent variable model for classification. In *Proceedings of the 24th international conference on Machine learning*, pp. 927–934, 2007.

Laurens van der Maaten. Learning a parametric embedding by preserving local structure. In David van Dyk and Max Welling (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 384–391, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL `https://proceedings.mlr.press/v5/maaten09a.html`.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

Qiaoya Wu and Yue Shen. A Catalog of Quasar Properties from Sloan Digital Sky Survey Data Release 16. *The Astrophysical Journal Supplement Series*, 263(2):42, December 2022. doi: 10.3847/1538-4365/ac9ead.

A The evidence lower bound $\mathcal{L}_x$

In this section we explicitly show the derivation of expected log-likelihood term in Eq. (6):

$$p(X) \geq \int p(F|U,Z)q(U)\log p(X|F,Z)dFdU - \mathrm{KL}(q(U)||p(U)) + \log p(Z) \tag{20}$$

$$\begin{aligned}\int p(F|U,Z)q(U)\log p(X|F,Z)dFdU &= \int q(\boldsymbol{u}_d)\int p(\boldsymbol{f}_d|\boldsymbol{u}_d,Z)\log\prod_{n=1}^{N}\prod_{d=1}^{D}\mathcal{N}(x_{n,d};f_d(\boldsymbol{z}_n),\sigma_x^2)d\boldsymbol{f}_d\boldsymbol{u}_d \\ &= \int q(\boldsymbol{u}_d)\int p(\boldsymbol{f}_d|\boldsymbol{u}_d,Z)\sum_{n,d}\log\mathcal{N}(x_{n,d};f_d(\boldsymbol{z}_n),\sigma_x^2)d\boldsymbol{f}_d\boldsymbol{u}_d \\ &= \sum_{n,d}\langle\log p(x_{n,d}|\boldsymbol{f}_d,\boldsymbol{z}_n,\sigma_x^2)\rangle_{q(.)}\end{aligned} \tag{21}$$

A.1 Tractability of the expectation term

The expectation term above is not just factorisable but also tractable, we show this explicitly below:

$$\begin{aligned}&\int q(\boldsymbol{u}_d)\int p(\boldsymbol{f}_d|\boldsymbol{u}_d,Z)\sum_{n,d}\log\mathcal{N}(x_{n,d};f_d(\boldsymbol{z}_n),\sigma_x^2)d\boldsymbol{f}_d\boldsymbol{u}_d \\ &\qquad = \int q(\boldsymbol{u}_d)\sum_{n,d}\log\mathcal{N}(x_{n,d};k_n^TK_{mm}^{-1}\boldsymbol{u}_d,\sigma_x^2) - \frac{1}{2\sigma_x^2}q_{n,n}\end{aligned} \tag{22}$$

$x_{n,d}$ is a scalar (d^{th} dimension of point $\boldsymbol{x}_n$), k_n^T is a $1\times M$ matrix - the n^{th} row of K_{nm}, we know that $p(\boldsymbol{f}_d|\boldsymbol{u}_d,Z) = \mathcal{N}(K_{nm}K_{mm}^{-1}\boldsymbol{u}_d, K_{nn} - K_{nm}K_{mm}^{-1}K_{mn})$. Further, $f_d(\boldsymbol{z}_n)$ is a scalar, denoting the value at index $\boldsymbol{z}_n$ of the vector $\boldsymbol{f}_d$. $q_{n,n}$ is the n^{th} entry in the diagonal of matrix $Q_{nn} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$. Next, performing the integration w.r.t $q(\boldsymbol{u}_d) = \mathcal{N}(\boldsymbol{u}_d;\boldsymbol{m}_d,S_d)$ yields,

$$\begin{aligned}\int q(\boldsymbol{u}_d)\Big[\sum_{n,d}\log\mathcal{N}(x_{n,d};k_n^TK_{mm}^{-1}\boldsymbol{u}_d,\sigma_x^2) - \frac{1}{2\sigma_x^2}q_{n,n}\Big]d\boldsymbol{u}_d &= \sum_{n,d}\Big[\log\mathcal{N}(x_{n,d};k_n^TK_{mm}^{-1}\boldsymbol{m}_d,\sigma_x^2) - \frac{1}{2\sigma_x^2}q_{n,n} \\ &\qquad - \frac{1}{2\sigma_x^2}\mathrm{Tr}(S_d\Lambda_n)\Big]\end{aligned} \tag{23}$$

Bringing it all together, the final lower bound can be written out explicitly as,

$$\begin{aligned}\mathcal{L}_x &= \sum_{n,d}\langle\log p(x_{n,d}|\boldsymbol{f}_d,\boldsymbol{z}_n,\sigma_x^2)\rangle_{q(.)} - \mathrm{KL}(q(U)||p(U)) + \log p(Z) \\ &= \sum_{n,d}\Big[\log\mathcal{N}(x_{n,d};k_n^TK_{mm}^{-1}\boldsymbol{m}_d,\sigma_x^2) - \frac{1}{2\sigma_x^2}q_{n,n} - \frac{1}{2\sigma_x^2}\mathrm{Tr}(S_d\Lambda_n)\Big] - \sum_d \mathrm{KL}(q(\boldsymbol{u}_d)||p(\boldsymbol{u}_d)) \\ &\qquad + \sum_n \log p(\boldsymbol{z}_n)\end{aligned} \tag{24}$$

where all the data dependent terms factorise enabling mini-batching of gradients. The shared model uses a sum of ELBOs $\mathcal{L}_x + \mathcal{L}_y$ with a shared latent embedding Z, which we call the joint evidence lower bound (Eq. (16) in the main paper). The additive structure ensures that the joint ELBO is also factorisable across data points N.

B Limitations

The primary limitation of the model arises from the fact that GPLVMs are decoder only models, they constitute a (potentially) smooth mapping from the latent space to the data space. This means that points close in latent space will be close in data space but not the other way round. Data space similarities can be preserved by including a back-constraint or an encoder which additionally maps from the data to latent space (Lawrence & Quiñonero Candela, 2006; Bui & Turner, 2015). The encoder usually takes the form of a neural network but other choices are possible. As we emphasized earlier, the model we propose is a decoder only model. The absence of an encoder complicates test-time inference as there is no deterministic way to access the latent points $\boldsymbol{z}^*$ corresponding to the new unseen observation $(\boldsymbol{x}^*, \boldsymbol{y}^*)$. Inferring the latent point corresponding to the unseen test point entails freezing the model and variational parameters post training and re-optimizing the ELBO objective to learn $\boldsymbol{z}^*$ with the new data point(s) included (see Algorithm 1).

This re-optimisation procedure for test-time inference ends up being too inconvenient in contrast to encoder-decoder models like VAEs (Kingma & Welling, 2013) where inferring a latent point corresponding to an unseen data point entails a forward pass through the trained encoder network (constant time predictions $\mathcal{O}(1)$). The main reason an extension to an auto-encoded shared stochastic GPLVM is not straightforward is due to the presence of missing data. The encoder network needs to be capable of handling arbitrarily missing dimensions. Methods like the partial VAE (Ma et al., 2018a;b) address this challenge in the context of VAEs where each observation is augmented with a column index indicating the observed dimension; the encoder network then processes these tuples as a 'set' with a permutation invariant set encoding function. A similar approach can be straightforwardly adapted our shared latent space set-up with Gaussian process decoders. The combination of a set encoder (to process missing dimensions) and a non-parametric decoder has not appeared in literature to the best of our knowledge. We are currently working on incorporating this feature into our framework. Note that the presence of an encoder acts as a constraint in the model, there is an inherent trade-off in terms of faster test inference and marginally weaker reconstructions / predictions.

C Experimental set-up

In this section we detail the configuration of the experiments in Section 7 of the main paper. For each of the datasets we repeat every experiment with 5 random seeds yielding different splits of the training data. The attributes of the data and sparse GP set-up are given in Table 3. We used a learning rate of 0.001 across all

Dataset	N	D	L	Num inducing M	Latent dim. Q
1K	996	590	4	120	10
22K	22844	657	4	250	10

Table 3: Experimental configuration to reproduce experiments in section 3 of the main paper.

parameters and ran the mini-batch loop with a batch size of 128 for 10,000 iterations on an Intel Core i7 processor with a GeForce RTX 3070 GPU with 8GB RAM memory. In order to give an estimate of the scale of the model for the 22k dataset we enclose a summary snapshot of the number of trainable parameters in our shared model.

D Cross-validating M and Q

The two main parameters of our shared framework which need to be fixed at the outset are: the number of inducing points M and the latent space dimensionality Q. We set M to be 250 in all the 22k experiments after cross-validation upto $M = 1000$ and found that $M = 250$ gave the best possible trade-off in terms of speed and accuracy. The reconstruction results with $M = 1000$ were only marginally better than with $M = 250$ inducing points but significantly increased compute due to the cubic scaling in inducing points.

In Fig. 8 we visualise the evolution of the ELBOs across varying latent dimensionality. We notice a meaningful improvement in increasing the dimensionality from $Q = 2$ but very marginal gains beyond $Q = 10$; we use this setting in experiments. It may be important to highlight that due to automatic relevance determination of

Modules	Parameters
inducing_inputs	2500
Z.Z	208440
model_spectra.variational_strategy._variational_distribution.variational_mean	164250
model_spectra.variational_strategy._variational_distribution.chol_variational_covar	41062500
model_spectra.mean_module.constant	657
model_spectra.covar_module.raw_outputscale	1
model_spectra.covar_module.base_kernel.raw_lengthscale	10
model_labels.variational_strategy._variational_distribution.variational_mean	1000
model_labels.variational_strategy._variational_distribution.chol_variational_covar	250000
model_labels.mean_module.constant	4
model_labels.covar_module.raw_outputscale	1
model_labels.covar_module.base_kernel.raw_lengthscale	10

Total Trainable Params: 41689373

Figure 7: Number of trainable parameters for the 22k model where `model_spectra` refers to the GPs corresponding to X and `model_labels` refers to the GPs corresponding to Y observation space.

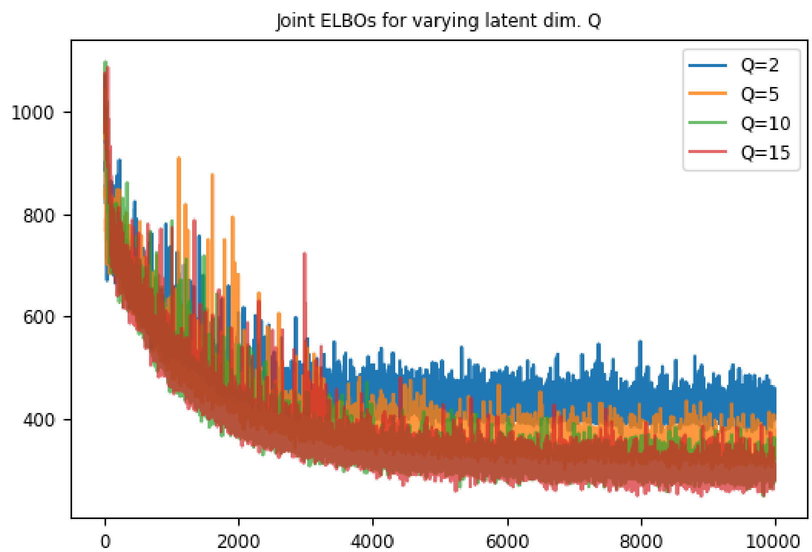

Figure 8: Sensitivity to Q: The negative ELBO objective for varying latent space dimensionality (lower is better)

the squared exponential kernel, setting a high latent dimensionality should not degrade results as the model automatically prunes redundant dimensions by driving the corresponding inverse lengthscales to 0. However, they do increase the compute cost, hence, it is important to set Q at a reasonable value which is flexible enough for structure discovery and not too constrained, while simultaneously minimising the computational burden.