Physics-Informed Gaussian Process Regression for Optical Fiber Communication Systems

Josh W. Nevin, *Student Member, IEEE*, F. J. Vaquero-Caballero, David. J. Ives, and Seb J. Savory, *Fellow, IEEE*

***Abstract*—We present a framework for enhancing Gaussian process regression machine learning models with a priori knowledge derived from models of the transmission physics in optical networks. This is done by framing the regression problem as multi-task learning, in which both the measured data and targets derived from a physical model of the system are used to optimise the kernel hyperparameters. We discuss the theoretical assumptions made and the validity of the approach. It is demonstrated that physics informed Gaussian processes facilitate Bayesian inference with fewer data points than standard Gaussian processes, opening up application areas in which measurements are expensive. The transparency, interpretability and explainability of the proposed technique and the subsequent increased likelihood of adoption by industry are discussed.**

***Index Terms*—Optical fiber communication, Gaussian processes, explainable machine learning, data-centric engineering.**

I. Introduction

MACHINE learning approaches have been applied to a wide range of problems in optical fiber communication networks [1], [2]. Many of these problems, such as quality of transmission (QoT) estimation, have also been approached by designing models based on the physics of optical fiber networks. These physical models, such as the widely-used Gaussian noise (GN) model [3] and split-step Fourier method (SSFM) [4], range in computational complexity and accuracy. In many cases, these physical models are not utilised but rather replaced by machine learning methods that learn to solve the problem directly from data. In this work, we present physics-informed Gaussian process (GP) regression, a methodology for the embedding of physical models within probabilistic machine learning.

Specifically, we address the following problems in this paper. Firstly, we wish to address the lack of machine learning approaches with a well-quantified predictive uncertainty which utilise the information from physical models that is known before we have taken any system measurements. To that end, we present a method for embedding physical models in GP regression, producing a physics-informed machine learning approach with a well-quantified predictive uncertainty. Knowing the uncertainty of model predictions is crucial within optical fiber communication networks, as these networks are typically established with a high availability [5] and thus model errors can result in catastrophic events, such as outages. Moreover, the proposed method addresses the problem that, due to uncertainties in the inputs to physical models, the predictions of such models may be inaccurate [6], even for highly complex models. To rectify this, the physics-informed GP method proposed allows us to update our physical model-derived estimate of the target signal with measurements of the signal to obtain an accurate GP predictive model. Another problem addressed in this work is that, as networks become increasingly dynamic, meaning that lightpaths are established and torn down with greater frequency, the volume of data available through network monitors becomes increasingly constrained. The proposed methodology addresses this problem by allowing us to train GP models with fewer measurements of the system, through the inclusion of information from physical models. Also, many of the machine learning approaches deployed within optical networks are black box methods, for which the decision processes within the algorithms are not transparent and the model predictions are not interpretable. Both of these factors mean that industrial trust in machine learning systems is often low, forming a barrier to deployment [7]. We address this by highlighting how the proposed physics-informed GP methodology is explainable, where in this work we follow the definition of explainability given by Rosher et al. [8]. In short, explainable machine learning algorithms are interpretable, meaning that the decisions made are human-understandable, transparent, meaning that the algorithm design is clearly motivated and include domain knowledge, meaning all the information we have about the target signal before we take any measurements of the system.

Therefore, the key contribution of this paper is the proposed physics-informed GP methodology that has a well-defined uncertainty and can be trained with fewer system measurements than standard GPs, due to the inclusion of information from physical models. We demonstrate this approach for a simple experimental system below, in order to explain the methodology, demonstrate its benefits and motivate further related study.

The rest of the paper is organised as follows. In Section II we highlight and briefly discuss related works from the literature, providing the context for our contribution. Following this, in Section III we outline the theoretical approach to integrating knowledge obtained from physical models with GPs and discuss the key assumptions made. We then describe the experimental system and corresponding physical model in Section IV, before presenting a demonstration of the benefits of the proposed methodology over conventional GP models in Section V-A and an exploration of key practical considerations for using this method in Section V-B. Furthermore, the explainability of the proposed method is discussed in detail in

Josh W. Nevin, F. J. Vaquero-Caballero, David J Ives and Seb J. Savory are with the Department of Engineering, University of Cambridge, Cambridge, UK, e-mail: jn399@cam.ac.uk.

Manuscript received Month Day, 2021; revised Month Day, Year.

Section V-C and concluding remarks are given in Section VI.

II. Related work

GP regression is a non-parametric, probabilistic machine learning technique for solving regression-style problems [9]. A detailed explanation of the theory of GPs is given in the book by Rasmussen and Williams [10], which has been used extensively in this work. Previous uses of GP regression in optical fiber communication networks include the work of Meng et al. [11], in which GP regression was used to predict the values of the signal to noise ratio (SNR) as a function of the transmission wavelength. Similarly, Wass et al. [12] used GPs to predict bit error rate as a function of launch power for an experimental WDM system. GPs can also be formulated to deal with classification problems in a probabilistic way. For example, Panayiotou et al. [13] trained GP classifier models on historical network data in order to determine the probability of failure for each network link.

A key theme of this paper is the integration of physical models with machine learning. An example of an approach that utilises both physical models and machine learning is that of Seve et al. [14], who presented a simple learning process that incorporated models of the physics of transmission in order to combat optical network design margins by improving the accuracy of the QoT estimation tool used in planning. Moreover, Seve et al. point out that physical models are imperfect and thus measurements should be used to refine these physical models to improve the quality of predictions, providing motivation for the work presented in this paper. Furthermore, Zhuge et al. [15] present a methodology in which neural networks (NNs) are combined with physical models for nonlinearity estimation. Here, the NN has two uses: to refine the errors of the physical model and to unify modelling and monitoring for nonlinearity estimation. Additionally, Raissi et al. [16] presented a methodology for the embedding of physical laws, represented by partial differential equations, within NNs. These physics-informed NNs have recently been applied within the optical networking domain for the first time [17], for the simplistic case of solving the nonlinear Schrödinger equation in an optical fiber. In this work we present an alternate approach to physics-informed machine learning, in which a GP machine learning method can be informed by physical models, allowing one to benefit from a well-quantified uncertainty level and enhanced explainability.

III. Theoretical approach

A. Standard Gaussian process regression

A GP is defined as a collection of random variables, any of which have a joint Gaussian distribution [10]. This Gaussian assumption facilitates analytical Bayesian inference, making GPs a powerful machine learning tool in which the level of uncertainty associated with predictions is quantified in a rigorous way. This uncertainty is easily interpretable, as the Gaussian assumption allows us to define confidence regions in terms of the number of standard deviations of variation away from the predictive mean, a metric that is easily understood. Furthermore, GPs are kernel-based methods, where a chosen kernel function is used to model the relationship between the data, facilitating more efficient learning via the kernel trick - working in feature space is possible because the algorithm is defined in terms of inner products in the input space [10][18]. Choosing a particular kernel means making assumptions about how we expect the data to vary and this choice should be made on a problem-by-problem basis. In this work, we choose the squared exponential plus a white noise kernel function, described by [10]

$$k(x_i, x_j) = h_1^2 \exp\left(\frac{-||x_i - x_j||^2}{2h_2^2}\right) + W(x_i, x_j) \quad (1)$$

where x_i and x_j are the input values of the points being compared, h_1 and h_2 are the hyperparameters of the squared exponential kernel, $W(x_i, x_j) = h_3^2$ if $x_i = x_j$ and 0 otherwise and $||\cdot||$ represents the Euclidean distance. Throughout the rest of this paper, we denote the set of kernel hyperparameters by $\theta = \{h_1, h_2, h_3\}$. By selecting this kernel, we assume that the data has one underlying length scale, controlled by h_2, with an absolute scale factor h_1 and independent and identically distributed Gaussian noise with variance ${h_3}^2$. We justify this choice of kernel in detail in Section V-C and remark that the method presented in this paper is general and is valid for any valid kernel function. Examples of situations requiring a more complex kernel include for signals which are expected to contain periodicity or those for which we expect the presence of multiple length scales. In order to design a kernel to use for a specific problem, one can utilise the fact that the sum of any two valid kernel functions is itself a valid kernel function to tailor the kernel to the individual problem being considered [10].

Furthermore, GPs are an example of a Bayesian approach to machine learning, meaning that Bayes theorem is used to generate a predictive probability distribution, the posterior, that combines the prior assumptions made about the problem with the measured data. In non-parametric methods such as GPs, the space of functions f is searched to find a functional model for the signal, rather than searching over a set of variable weights for a fixed functional form such as in a parametric model. We can write Bayes rule in the context of GPs as [10]:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} = \frac{p(y|f, X, \theta)p(f|X, \theta)}{p(y|X, \theta)}. \quad (2)$$

Here, the prior distribution contains assumptions about how we expect the data to vary, the likelihood is the probability of the data targets given the data inputs X and kernel hyperparameters θ, and the marginal likelihood can be thought of as a normalisation factor.

In order to fit a GP model to a dataset, we need to first optimise the values of the kernel hyperparameters. This is done by maximising the log marginal likelihood - the probability of the data targets given the inputs, marginalised over the space of functions - using a gradient-based method, as outlined in [10], Chapter 5. Thus, we aim to search the space of functions to find the most likely interpretation of the data. The log marginal likelihood and its gradient can be calculated as [10]

$$\log p(y|X, \theta) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi \quad (3)$$

$$\frac{\partial}{\partial\theta_j}\log p(y|X,\theta) = -\frac{1}{2}\mathrm{Tr}\Big((\alpha\alpha^T - K^{-1})\frac{\partial K}{\partial\theta_j}\Big), \quad (4)$$

where K is the kernel matrix consisting of $k(x_i, x_j)$ for all i, j, n is the number of data points, y are the data targets, X is the matrix of input data and $\alpha = K^{-1}y$. Once optimal hyperparameters have been found, Algorithm 2.1 from Rasmussen and Williams [10] is used to fit the GP to a given dataset, by calculating the predictive mean and variance of the model [10]:

$L = \mathrm{cholesky}(K + \tau I)$ - Cholesky decomposition
$\alpha = L^T \setminus (L \setminus y)$
$\bar{f}_* = k_*^T\alpha$ - GP predictive mean
$v = L \setminus k_*$
$V[f_*] = k(x_*, x_*) - v^T v$ - GP predictive variance

where here $\bar{f}_*$ and k_* refer to the predictive mean and the covariance function evaluated at the test point x_* respectively, τ is a small constant added to the diagonal of K to ensure that the calculated matrix is positive-definite [19] and I is an $n \times n$ identity matrix. The value of τ used here is the default for the Scikit Learn GP library of 10^{-10} [19], which was used to fit the GP models in this work. More specifically, the Scikit Learn GP models optimise the hyperparameters by gradient-based maximisation of the log marginal likelihood using the SciPy implementation of the L-BGFS-B algorithm [20], [21]. This algorithm is called with a default maximum number of iterations and function evaluations of 15,000, which was not met by any models presented in this work, as well as two stopping criteria. The first criterion is defined by the normalised difference between function evaluations at successive steps, for which the default value of the order of 10^{-9} is used. If we have reached a local optimum, the gradient around the optimum will be small, and thus these differences should be small. The second criterion is defined by a lower bound on the maximum element of the projected gradient, for which again the Scikit Learn default threshold of the order of 10^{-5} was used. Both of these criteria are designed to induce stopping of the gradient-based method at a local optimum, where the gradient is sufficiently small. In order to increase the probability of finding the global optimum, the optimiser is restarted 20 times and run from different randomly selected initial conditions each time, within the broad hyperparameter bounds given. It should be noted that none of the GP models presented returned optimal hyperparameters equal to these bounds. The resulting model with a maximum log marginal likelihood is then selected. It is also important to note that we make the distinction between hyperparameter optimisation and the process of computing the predictive mean and variance given a set of optimal hyperparameters, referring to the latter as fitting the GP.

B. Proposed method for including physical models

We propose to include physical models in the optimisation of the kernel hyperparameters of the GP regression model, such that our a priori knowledge of the system can be incorporated. In this work, we define our a priori knowledge as all the information that we have about the target signal before we have made any measurements of the system. This definition is synonymous with the definition of domain knowledge given in Rosher et al. [8]. As discussed in detail in Section V-C below, a priori knowledge can take a number of forms and is difficult to define in general. The specific target signal in this work is the SNR as a function of the launch power and our a priori knowledge is made up of the following components:

1) The physics-based model of the target signal, which provides an approximation to this signal for a set of parameters with a given uncertainty, as well as describing the general behaviour of the system as a function of the target input.
2) A set of estimated system parameters, each of which with an estimated degree of uncertainty. These uncertainties may be bounded by specifications given in equipment data sheets.
3) All context provided by the relevant literature.

Our goal is to embed this a priori knowledge into a GP regression model, such that our approximate knowledge of the signal can be updated with measurements of the system. Mathematically, we can express the a priori knowledge given by the physical model as a simple formula, as is given below in Section IV. This physical model can also be used to make decisions about the range of measurements that we should make. For example, if we can estimate the launch power that gives an optimal SNR from the physical model, we can make an informed decision about the range of launch powers over which we take measurements. Similarly, the parameters of this model can be represented as random variables, each drawn from some unknown distribution. These parameters relate to a given physical component of the system, such as the attenuation coefficient of the optical fiber used. Bounds for these parameters are often given by vendor specification sheets. Ultimately, some estimate of the parameters is used for the physical model, producing a set of approximate predictions of the target signal. It is more difficult to express the contribution to the a priori knowledge from the literature, as this will be highly problem-specific. An example may include measurements reported for a similar system, that can be used to obtain an estimate for the system parameters or the target signal of interest.

The key approach taken for the inclusion of a priori knowledge within GPs in this work is to frame the problem as an example of multi-task learning. In multi-task learning, the goal is to find one common optimal set of hyperparameters for multiple tasks simultaneously, such that these hyperparameters produce a model that benefits from similarities and differences across the tasks [10], [22]. Once the optimal hyperparameters have been found for the two tasks, we use them to fit a model that performs one of the tasks more effectively than if only this single task was considered.

We propose to use an a priori physical model of the system to generate a set of targets and to perform multi-task learning, using these physical model-generated targets and the measured data in the optimisation of the hyperparameters. Here the multi-task learning framework consists of two regression tasks - one for the physical model predictions and one for the measured data. The task of interest to us is to perform regression on the measured data, and performing this regression with hyperparameters optimised for both tasks allows us to embed information from the physical model our GP regression model. When we use multi-task learning in this way, we assume that the physical model targets and the measured data are described by the same underlying statistical distribution. Specifically, this means assuming that the targets in both datasets are independent and identically distributed (i.i.d.) random variables drawn from the same distribution. The i.i.d. assumption is widespread across statistical learning theory and underpins the majority of modern machine learning algorithms [23]. Thus, we must take care to ensure that this assumption is reasonable. Physical models are commonly deterministic, with no element of uncertainty included. To facilitate multi-task learning, we include uncertainty in these models in a physical way, as outlined in Section IV.

More formally, in order to perform multi-task learning, we generate physical model targets y_p for input parameters X_p, analogous to the measured data targets and input parameters y and X, and find kernel hyperparameters θ that maximise the sum of the log marginal likelihoods for both sets of targets:

$$\arg\max_{\theta}[\log p(y|X,\theta) + \log p(y_p|X_p,\theta)] \tag{5}$$

where the log marginal likelihoods and their gradients are calculated using (3) and (4) respectively. It should be noted that the physical model targets can be generated at different X values to the data, hence the specification of X_p. This also means that we can use a different number of physical model targets to the number of data targets. We investigate the effect of changing the number of physical model targets in Section V-B. Then, the GP is fitted to the data in the standard way using Algorithm 2.1 from Rasmussen and Williams [10], reproduced above, using only the data targets. Taking this multi-task learning approach allows for the information in the physical model to be taken into account when optimising the kernel hyperparameters, meaning that we can include our a priori knowledge of the system in the model.

In standard GP models, the computational complexity is dominated by the inversion of the K matrix, which scales with the number of data points n as $O(n^3)$. When performing multi-task learning, if the number of physical model targets n_p is different to that of the data targets, a different K must be calculated and thus the same inversion must be computed for the physical model-generated targets, which also scales with the number of physical model targets as $O(n_p^3)$. If $X = X_p$ however, then the same K is used in both cases and the computational complexity is $O(n^3)$. Thus, for $X \neq X_p$, for which n and n_p can differ, the dominant term depends on the relative size of n and n_p. Choosing a suitable value of n_p is discussed further in Section V-B.

Moreover, we note that this method is not equivalent to using physical models to generate synthetic data points and using these both for hyperparameter optimisation and calculation of the model predictive mean and variance. By only using the physical model-generated targets in the optimisation of the hyperparameters, we are improving our estimates of the hyperparameters by imparting our a priori knowledge of the system via multi-task learning. This is an important distinction to make, as using the physical model targets as synthetic data that is treated in the same way as measured data would yield models that are very sensitive to errors in the physical model. This methodology has been implemented as a modification to the Scikit-learn Gaussian Process Regressor class [19], in which new methods have been added to perform multi-task learning.

IV. Example physical system

A. Experimental setup

In this work we use measurements of a simple point-to-point link, originally presented in [24], consisting of 10 spans of Corning SMF-28 fibre, each of length 100 km. This system is depicted in Figure 1. A Polatis 32×32 Fiber Switch is used to connect the individual 100 km spans together, which themselves consist of fiber spools contained within a rack. This switch is also used to control the launch power into the spans, which in this experiment was uniform across the spans. A 25 dB fixed gain Erbium-doped fiber amplifier (EDFA) is used, along with a variable optical attenuator (VOA) to compensate for the extra gain. A Ciena WaveLogic 2 coherent transceiver is used to transmit a single channel at 11.5 GBaud using a dual-polarisation quadrature phase-shift keying (QPSK) modulation format. The signal is measured by an optical spectrum analyser after each amplification, which is used to determine the attenuation required to set the desired launch power into each span. The output of the final span is then passed through an optical channel filter in order to remove any amplified spontaneous emission noise beyond the channel bandwidth. The signal was then received and recovered by the WaveLogic 2 transceiver, following which the SNR was estimated from the received constellation via the technique proposed in [25], in which the radial moments of the QPSK signal constellation are used to perform the estimation. This simple system is used to demonstrate the proposed physics-informed GP methodology and explore the key practical considerations necessary to deploy this technique. As discussed above, it is the technique for inclusion of physical model information in GP models that is the focus in this work. The technique proposed is applicable to higher-dimensional input spaces and for signals with more complex features, such as periodicity and the presence of multiple length scales. Specifically, we used this experimental system to generate a dataset of SNR as a function of the launch power into each span for a single channel. The input power was uniform across the 10 spans, creating a simple, well-understood dataset of SNR as a function of the uniform launch power.

B. Physical models

We use a physical model in order to generate an approximation to the signal based on our a priori knowledge of the experimental system, as defined in Section IV, given by [26]

$$\mathrm{SNR_{phys}} = \left(\frac{a + bP_{in}^3}{P_{in}} + \frac{1}{\mathrm{SNR_{TRx}}}\right)^{-1}, \tag{6}$$

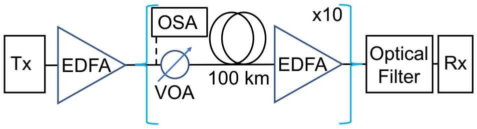


Fig. 1. Simple point-to-point system used to generate the demonstrative dataset of SNR as a function of launch power, consisting of 10 spans of length 100 km, amplified using a 25 dB fixed gain EDFA. A single QPSK channel is transmitted and the SNR is estimated from the recovered signal at the receiver using the radial moments of the QPSK signal.

where the constants a and b represent the strength of the linear and nonlinear noise respectively, $\mathrm{SNR}_{\mathrm{TRx}}$ is the back-to-back SNR of the transceiver and P_{in} is the launch power. Furthermore, as discussed in Section III, in order to perform multi-task learning, the datasets used are assumed to be drawn from the same statistical distribution. In order to satisfy this assumption, the physical model must not be deterministic, and should capture the uncertainty in the system. There are two types of input into the physical model - the input variables X, namely the launch power in this case, and the parameters that represent the physical characteristics of the system, such as the fiber loss and fiber nonlinearity coefficient. Both types of input have an associated uncertainty, however the launch power is a variable which is changed throughout the experiment. Therefore, we include the launch power uncertainty in the physical model because the launch power is a stochastic random variable. Contrastingly, the physical layer parameters, such as the fiber loss and nonlinearity coefficient, are assumed to be constant over the duration of the experiment. As a result, the uncertainty in these model parameters is not modelled within the physical model itself. In a deployed system, some of these parameters will change with time, however the scale with which they change will be long relative to the time taken to make measurements as they are due to processes such as fiber ageing which have a slow rate of change [27]. Moreover, for lightpaths that are established over long periods of time, changes in these parameters can be accounted for by re-measuring the system and re-training the GP, providing an updated GP model.

Specifically, the Polatis 32×32 switch specification sheet quotes a maximum error of ±0.5 dBm. Thus, we model the launch power values as being Gaussian-distributed, with a mean equal to the measured value P_{meas} and standard deviation $\sigma = 0.5/3 = 0.167$ dBm, such that 99.7% of the perturbation values lie within ± 0.5 dBm. Thus, we model P_{in} as

$$P_{in} \sim \mathcal{N}(P_{meas}, \sigma^2). \tag{7}$$

We then use the SSFM to obtain a priori estimates for the parameters a and b. There exist alternative, more approximate physical models to the SSFM that are less computationally intensive, such as the GN model. The SSFM has been chosen over such models in this case primarily because of the experimental system considered, which has parameters for which the signal Gaussianity assumption used in the GN model is a poor approximation to the signal. Namely, the system has a low symbol rate, a single channel is transmitted and a QPSK modulation format is used, all of which mean that the signal is far from being Gaussian-distributed, as assumed by the GN model. It should be noted that there are corrections to the GN model which relax this Gaussianity assumption, including the enhanced GN (EGN) model [28]. However, despite the existence of proposed analytical approximations to the EGN [29] there is a lack of clear consensus on which closed-form formula to use, and thus using the EGN would involve numerical computation of integrals. Therefore, we decide to use the SSFM due to its superior accuracy and assume that the physical model targets can be computed offline. Thus, it is important to note that in this work, the majority of the uncertainty in the physical model is due to uncertainty in the inputs to the SSFM model, which are significant for deployed systems, as discussed in [6]. Therefore, here we aim to use an accurate physical model with uncertain inputs to generate our a priori estimate of the signal, before updating this estimate with measured data to produce an accurate predictive GP model.

Specifically, to obtain the parameter estimates from the SSFM, we use initial estimates for the parameters of the system, obtained from the specification sheets of the equipment and the literature, to generate a set of simulated SNR values. We generate SSFM simulations at launch power values of -4 dBm, 0 dBm, and 4 dBm, to ensure coverage of the SNR optimum, which we know will lie in this broad range a priori. From these simulations, we then fit (6) with a and b as free parameters via the method of least squares. Note that the SSFM does not include the effect of the back-to-back transceiver SNR, $\mathrm{SNR}_{\mathrm{TRx}}$, which is estimated from [30] to be 14.8 dB for the Wavelogic 2 linecard used. Based on our a priori physical model of the system, we calculate the range over which we measure the SNR by using the physical model to estimate the launch power corresponding to the optimum SNR, and then calculating the launch power values that correspond to a maximum of 2 dB SNR penalty in the linear and nonlinear regimes. This yields the range -8 dBm to 4 dBm to the precision of 1 dBm used in [24].

Using this a priori physical model, we can facilitate training GP models with sparse datasets, minimising the SNR penalty incurred when taking measurements and aiding with the development of online, data-driven networks. These applications are demonstrated in Section V-A below.

To achieve a transparent model, it is important to justify the chosen model inputs, which are outlined below.

- NLI coefficient, $\gamma = 1.2$ /W/km
- Dispersion coefficient, $D = 17$ ps $\mathrm{nm}^{-1}\mathrm{km}^{-1}$
- Loss, $\alpha = 0.2$ dB km^{-1}
- EDFA noise figure, $\mathrm{NF} = 4.6$ dB
- Operating wavelength, $\lambda = 1550$ nm
- Symbol rate, $R_s = 11.5$ GBd
- TRx back-to-back SNR, $\mathrm{SNR}_{\mathrm{TRx}} = 14.8$ dB
- P_{in} noise standard deviation, $\sigma = 0.167$ dBm

The values of D, α and NF are estimated from from the fibre and EDFA manufacturer specification sheets and γ is taken as an initial estimate for the Corning SMF-28 optical fiber used. λ, R_s and Δf are set when the experiments are performed and $\mathrm{SNR}_{\mathrm{TRx}}$ is estimated from [30]. σ is defined from the specification sheet of the 32×32 switch, as described above.

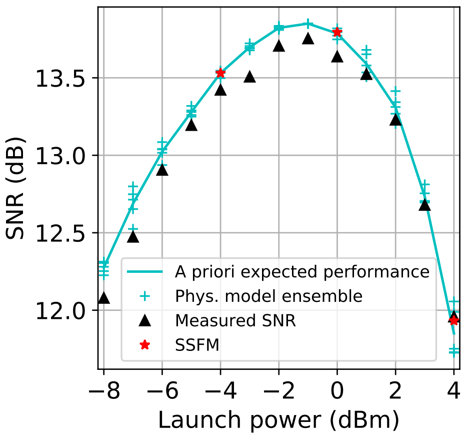


Fig. 2. An ensemble of 5 physical model predictions generated using (6), with $\sigma = 0.167$ dB and the a priori expected performance, computed as the ensemble mean, are compared with demonstrative dataset consisting of measured SNR as a function of launch power, generated using the experimental setup outlined in Figure 1. SSFM simulations used to optimise the parameters of the physical model are shown for comparison.

Due to the uncertainty present in the physical model, we draw an ensemble of 5 physical model predictions at each launch power value and compare to the measured SNR as a function of launch power from the experimental setup in Figure 1 and the SSFM simulations used to optimise a and b. This comparison is shown in Figure 2. Our a priori expected performance, computed as an average of the ensemble, is also shown. We wish to refine this a priori estimate of the signal variation by using measurements to obtain a more accurate posterior distribution with a quantified uncertainty level.

V. Demonstration of physical-model enhanced Gaussian processes

A. Compensating for sparse data

Here we demonstrate how including information from physical models in the optimisation of GP hyperparameters allows us to perform Bayesian inference with fewer measurements of the system. This would be of practical significance in any situations where the amount of available data is strongly constrained, such as when there is a performance penalty associated with making measurements or in a future dynamic optical network, where lightpaths are put up and torn down rapidly.

As an example, we use a sparse data subset of 5 measurements and 15 physical model targets, generated using (6), to fit a physics-informed GP. Figure 3 shows the physics-informed GP model, along with the measured data targets and physical model targets used in fitting. A standard GP fitted to the same dataset is also shown for comparison, as well as a two standard deviation confidence region, where the $x\sigma$ confidence upper and lower bounds are calculated using [10]

$$x\sigma \text{ confidence bounds} = \bar{f}_* \pm x\sqrt{V[f_*]}. \tag{8}$$

From Figure 3a, it can be seen that the physical model targets used in the hyperparameter fitting allow for the underlying signal to be estimated with only 5 measured data points with the physics-informed GP. Using the approach outlined in Section III, we have used the measured data and the a priori knowledge provided by the physical model and system parameter estimates during hyperparameter optimisation, in order to obtain a more accurate model of the system. From Figure 3b, it is clear that the standard GP has insufficient information to learn the signal variation without the inclusion of physical model targets. Furthermore, we also fit a standard GP to a larger dataset consisting of 13 measurements of the SNR made over the same range of launch power values, inclusive of the sparse dataset. This represents the case where we have more than the required number of measurements in the given domain to estimate the signal from measured data alone, providing a reference to compare to the physics-informed GP. In Figure 4, the predictive mean of this standard GP is compared to the predictive mean function of the physics-informed GP from Figure 3a, which is trained on the sparse dataset. To assess the accuracy of these models, we compare the mean absolute error (MAE) of the predictive mean with respect to the full dataset of 13 measurements. We find that this MAE is 0.02 dB higher for the physics-informed GP trained on the sparse dataset as compared to the standard GP trained on the full dataset, a relative increase of 3%, demonstrating that the physical model enables us to achieve comparable model accuracy in this case with 5 data points, rather than 13.

Additionally, we are able to train a physics-informed GP on a sparse subset of 3 measurements, taken at the two extremes and at the predicted SNR peak. This is the smallest number of measurements over the domain that can be used for inference. This model is shown in Figure 5 along with a standard GP fitted to the same sparse dataset for comparison. Note that as the GP is homoscedastic, the predictive variance $V[f_*]$ does not vary significantly with the launch power and hence we take the mean of the predictive standard deviation, $\overline{\sqrt{V[f_*]}}$, to compare different models. For the subset of 3 measurements, the predictive standard deviation of the physics-informed GP is much larger as compared to when 5 measurements are used, increasing by a factor of 4.7 from 0.13 dB to 0.61 dB, reflecting the fact that we have given the GP less information. Moreover, the predictive mean MAE with respect to the full dataset of 13 measurements is 0.06 dB higher as compared to the standard GP trained on 13 measurements, approximately 3 times greater than that achieved with 5 measurements. Thus, we are able to perform Bayesian inference with only 3 measurements, at the expense of some degree of model confidence and accuracy. For the standard GP, we find that a data subset of 6 or more measurements of the SNR is required for inference of the signal and show the physical model enhanced-GP and standard GP models trained on this subset in Figure 6. However, the mean of the predictive standard deviation is 0.12 dB for the physics-informed GP as compared to 0.19 dB for the standard GP, a relative decrease of 37%. This indicates that the inclusion of prior knowledge from the physical model results in a more confident GP model prediction, as the GP has more information regarding how the signal is expected to vary. For these models, the MAE with respect the full dataset of 13 measurements differs from that of a standard GP trained on the full dataset by only 0.01 dB and 0.02 dB for the physics-informed GP and standard GP respectively - these models are compared in Figure 7.

Thus, for this system, inclusion of a priori knowledge in GPs via physical models allows us to make up to 50% fewer measurements of the system when performing Bayesian inference, depending on the required accuracy and model confidence. Moreover, even with sufficient measurements for the standard GP to infer the signal, the physics-informed GPs are able to make more confident predictions, due to the a priori knowledge that has been provided during hyperparameter optimisation. This illustrates that, in general, the a priori knowledge of the system is less important in situations where we have a large amount of data. However, when the amount of data is constrained, our prior knowledge of the system becomes crucial, with the measured data points updating this prior to obtain a more accurate model. It should also be noted that the intention of the analysis in this section is to demonstrate the effect of including a priori knowledge in the GP using a simplistic, example dataset. This technique is general and applicable to more complex, higher-dimensional regression problems.

B. Practical considerations

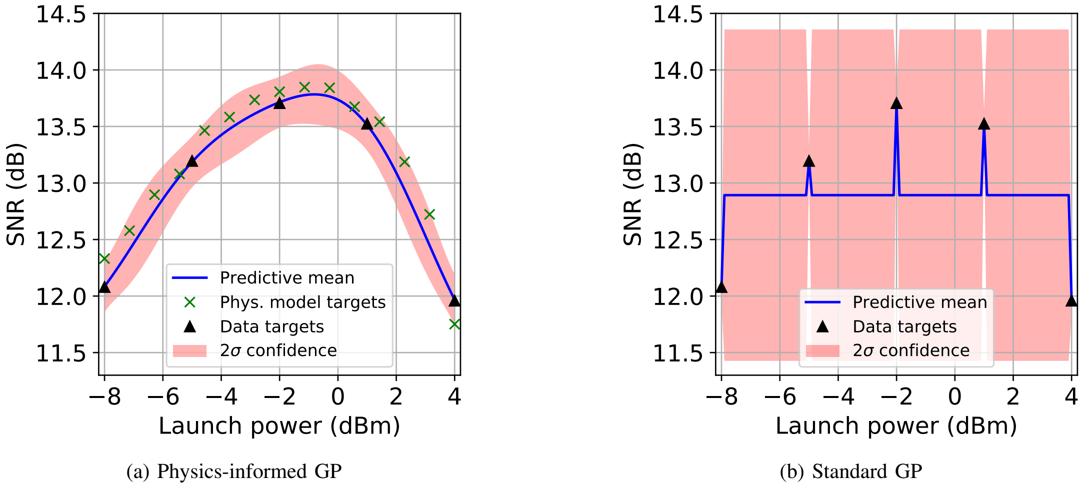
It is important to investigate the effect of changing the value of σ in (7) on the physics-informed GP, as this has been estimated from the bounds provided from the data sheet of the optical switch and thus itself has an associated uncertainty. Varying σ corresponds to modifying the variance of experimental noise that is added to the physical model, thus changing the uncertainty of the resulting GP, as the data targets and physical model targets are assumed to share the same optimal hyperparameters. Thus, physics-enhanced GP models are fitted to the sparse dataset of 5 SNR measurements, with 15 physical model targets used in the hyperparameter optimisation. The effect of changing σ on the kernel hyperparameters is demonstrated in Figure 8. We can see that the noise level hyperparameter, h_3, increases as we increase σ - as we tell the GP a priori that there is more noise in the system, the model reflects this through an increase the in hyperparameter controlling the noise. In response to the increased noise level, the absolute scale h_1 of the process decreases with increasing σ, whilst the length scale h_2 is relatively constant. Additionally, we consider the effect of σ on the mean of the predictive standard deviation of the GP model. Figure 9 demonstrates that as we increase the standard deviation of the added noise, we see a corresponding increase in the predictive standard deviation of the physics-informed GP. As we increase the a priori noise in the physical model, the resulting model makes predictions with a lower confidence. Moreover, we also remark that the MAE of the predictive mean function of the GP with respect to the full dataset of 13 measurements is constant with respect to σ to within 0.01 dB, highlighting that it is the predictive variance of the model that is primarily effected by the choice of σ, rather than the predictive mean.

It is also important to consider the effect of the number of physical model-generated targets on the physics-informed GP. For the example dataset used, there are 5 measured data points and we must choose how many physical model targets to use for hyperparameter optimisation. Thus, we vary the number of physical model targets used with a fixed added noise $\sigma = 0.167$ dBm and record the variation of the kernel the hyperparameters for the physics-informed GP in Figure 10. Firstly, we observe that for low numbers of physical model targets, approximately below 10, the kernel hyperparameters vary significantly, indicating that we have an insufficient number of physical model targets to learn the signal variation with only 5 measured data points. For 10 or more physical model targets however, the absolute scale h_1, length scale h_2 and noise h_3 are relatively constant with respect to the number of physical model targets. We conclude that the number of physical model targets used in the hyperparameter optimisation has no significant effect on the optimal hyperparameters found, provided a sufficient number of targets are used relative to the number of data points. It should be noted that if we have a higher density of measured data in the same range, then fewer physical model targets are required to learn the signal variation, as we can rely more on the contribution from the measured data to the log marginal likelihood in (5). Furthermore, the variation of the mean of the predictive standard deviation $\sqrt{V[f_*]}$ is reported in Figure 11 for the physics-informed GP, showing a sharp drop as we increase the number of physical model targets, followed by a plateau. Again, this indicates that with insufficient physical model targets for a given number of measurements, the model has a lower confidence in its predictions. Once a sufficient number of physical model targets are included, the model confidence plateaus and adding more targets has a minimal effect on the model confidence. Thus, we conclude that in general the choice of the number of physical model targets has a small effect on the resulting model compared to the value of σ from (6), provided a sufficient number of targets is used, in this case approximately 10 or more.

C. Explainability of proposed model

Recently there has been an increasing focus on the explainability of machine learning models, however the term explainability is very broad and often ill-defined. In the context of optical fiber communications, the use of explainable machine learning models is largely motivated by trust - within any industry, algorithms for which the decision processes can be understood by operators are more likely to be adopted [7]. Here we define the term explainability and highlight which areas of the proposed model are explainable. An explainable machine learning model can be defined as one that is transparent, interpretable and includes domain knowledge of the target problem [8]. Here a priori knowledge and domain knowledge represent the same concept, and we refer to a priori knowledge only in this section.

A transparent model is one for which the model designer can explain the reasoning behind the design choices made, beyond empirical success on the test data. To some degree,



(a) Physics-informed GP

(b) Standard GP

Fig. 3. Physics-informed GP and standard GP models trained on a sparse dataset of 5 measurements, with two standard deviation confidence region shown. For the physics-informed GP, $h_1 = 5.34$, $h_2 = 3.49$ and $h_3 = 0.0120$, whereas for the standard GP $h_1 = 0.996$, $h_2 = 1.83 \times 10^{-5}$ and $h_3 = 3.63 \times 10^{-3}$. The physical model targets used in hyperparameter fitting are shown as well as the data targets in (a). The standard GP has insufficient information to perform Bayesian inference of the signal.

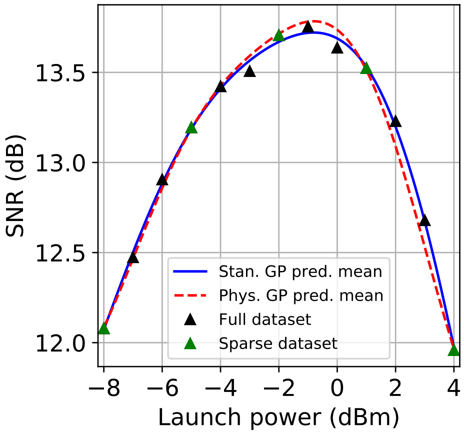
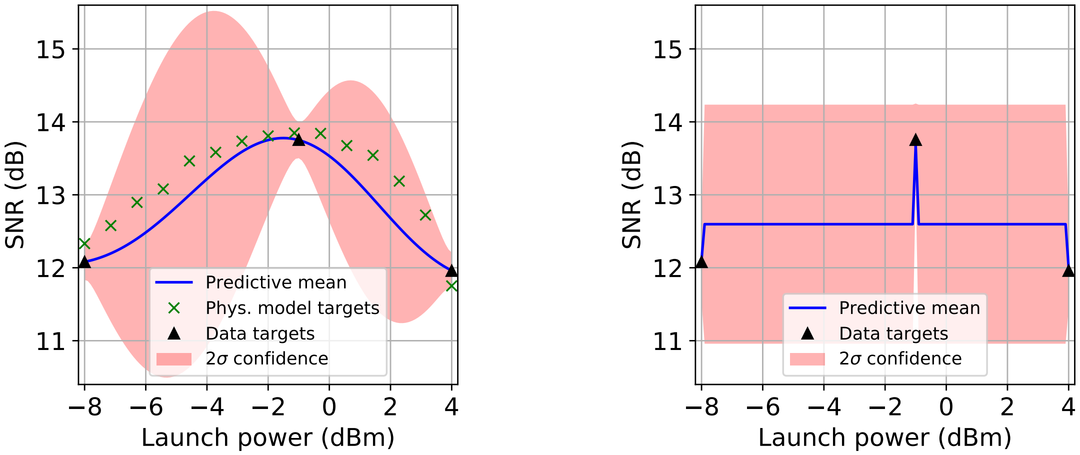


Fig. 4. Comparison of the predictive mean of the standard GP trained on the full dataset of 13 measurements and the predictive mean of the physics-informed GP, trained on the sparse dataset of 5 measurements.

transparency can be thought of as the opposite of a black-box [31]. An interpretable model is one for which the model output can be understood by a human, where an interpretation is defined as a translation of the model prediction to a human-understandable domain [32]. Interpretable models can be further sub-divided into two broad classes; those that are interpretable by design and those for which we accept that the model is a black box and seek to understand the input-output relationship only. The latter techniques are known as post-hoc, with widely used examples presented in [33], [34]. A priori knowledge is difficult to define in general and can loosely be thought of as all knowledge of the problem before we have seen the data [35]. In this work, we have outlined what we mean by a priori knowledge in Section III. An example of a method that does not include any a priori knowledge would be a black box model, such as a standard feedforward NN, which is trained blindly on a dataset to infer a relationship between some data X and some target data y. A priori knowledge can be structured to different degrees, for example it can be written as a mathematical formula, as in (6), or it can take the form of the bounds provided by the specification sheets of the equipment used in this work. Moreover, this a priori knowledge can be integrated with machine learning models in a number of different ways [8]. In this work we have chosen multi-task learning for instance, whereas a different method for inclusion was chosen in [14].

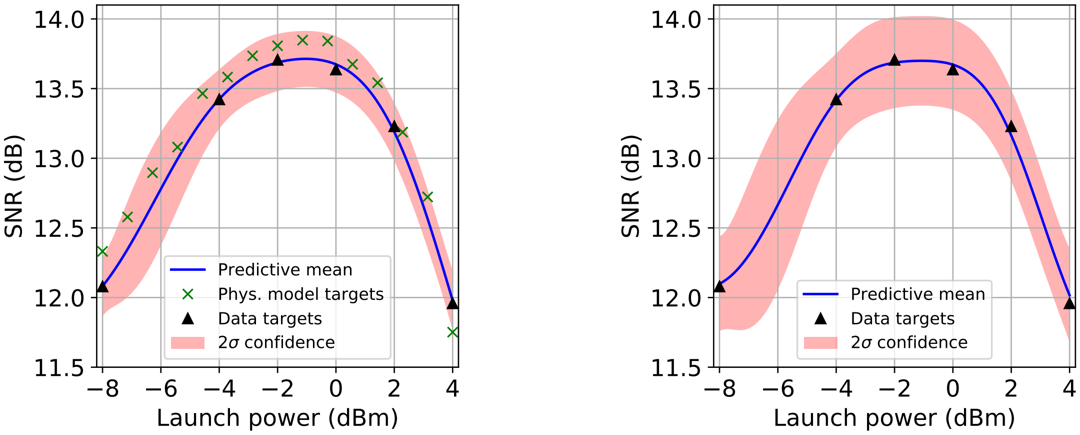
In general, kernel methods such as GPs can be called transparent, as the chosen kernel function contains the features that we expect to see in the target signal a priori. Moreover, the kernel function is composed of a sum of kernel functions, each representing a given set of features, which makes the model transparent [8]. In this work, our choice of kernel is transparent, as we know a priori that the signal, SNR as a function of launch power, can be described by a single length scale and contains no other features, such as periodicity or decay. Moreover, we justify the choice of white kernel by considering that the signal uncertainty is the result of a number of different experimental sources of uncertainty, each described by some statistical distribution, and thus can be approximated by a Gaussian via the central limit theorem. Furthermore, the multi-task learning methodology for including physical model-generated targets is itself transparent. This is because it can be justified and its success explained. We wish to embed knowledge from the physical model in a probabilistic machine learning method with well-quantified uncertainty, which is done by learning a set of optimal hyperparameters that describe both the physical model and the measured data. Specifically, we maximise the sum of the contributions to the log marginal likelihood from the measured data and physical model targets, hence imparting a priori knowledge in the kernel hyperparameters. We then fit a GP model to the measured data using these optimal hyperparameters. This allows us to do Bayesian inference with fewer measurements,



(a) Physics-informed GP

(b) Standard GP

Fig. 5. Physics-informed GP and standard GP models trained on a sparse dataset of 3 measurements, with two standard deviation confidence region shown. For the physics-informed GP, $h_1 = 4.42$, $h_2 = 3.11$ and $h_3 = 0.0115$, whereas for the standard GP $h_1 = 0.938$, $h_2 = 0.0185$ and $h_3 = 0.0623$. The physical model targets used in hyperparameter fitting are shown as well as the data targets in (a). The standard GP has insufficient information to perform Bayesian inference of the signal.



(a) Physics-informed GP

(b) Standard GP

Fig. 6. Physics-informed GP and standard GP models trained on a sparse dataset of 6 measurements, with two standard deviation confidence region shown. For the physics-informed GP, $h_1 = 7.20$, $h_2 = 3.75$ and $h_3 = 0.0114$, whereas for the standard GP $h_1 = 1.54$, $h_2 = 2.61$ and $h_3 = 0.0284$. The physical model targets used in hyperparameter fitting are shown as well as the data targets in (a). The inclusion of prior knowledge in the physics-informed GP leads to a 37% increase in prediction confidence over the standard GP.

as the physical model targets provide extra information about the signal during hyperparameter optimisation.

Furthermore, we consider the interpretability of the proposed method. Figures 8, 9, 10 and 11 demonstrate the interpretable nature of the physics-informed GP method, as we can make sense of the hyperparameter variation and the predictive variance of the model as we alter the physical model parameters. This partly relies on the principled nature of GPs, but is also due to the easily interpretable way in which the physical model targets have been included in the hyperparameter optimisation process. Furthermore, for regression problems more complex than the simple example considered here, analysis of the kernel hyperparameters can be used to understand the key features of the model and how they respond to variations in the input. For instance, if a periodic kernel function is used, the hyperparameter controlling the frequency of the periodicity can be used to easily interpret the periodicity in the signal. Thus, studying the hyperparameter variation is an effective method for providing interpretation of the GP output. Additionally, the confidence regions provided by GP models aid in the interpretation of predictions, as the degree of confidence provides extra information as compared to a model that simply makes predictions with no associated confidence. The human that is looking at the GP output can judge how confident the model is of its prediction, which can help motivate decisions such as whether or not to acquire more data and whether or not the model prediction can be trusted. Also, in this work a priori knowledge has been integrated within the machine learning methodology explicitly through multi-task learning. Furthermore, a priori knowledge has been

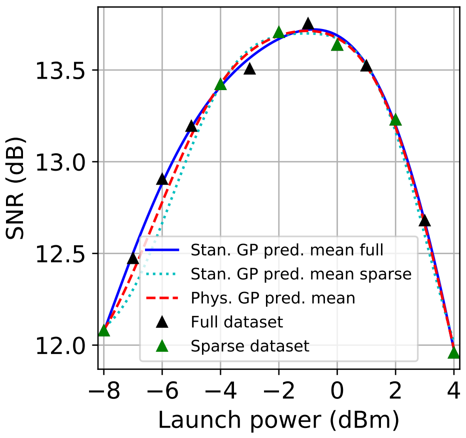


Fig. 7. Comparison of the predictive mean of the standard GP trained on the full dataset of 13 measurements and the predictive mean functions of the physics-informed GP and standard GP, both trained on a sparse dataset of 6 measurements.

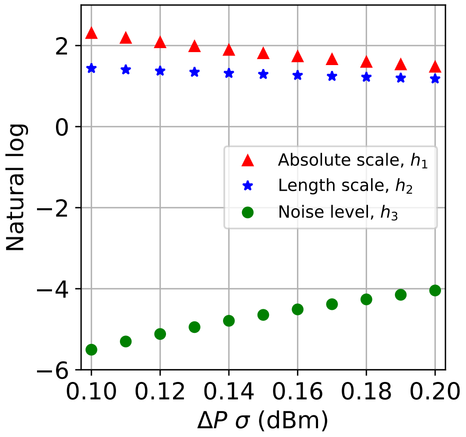


Fig. 8. Variation of kernel hyperparameters representing the absolute scale h_1, length scale h_2 and noise level h_3, defined in (1), with standard deviation of the launch power noise in (6).

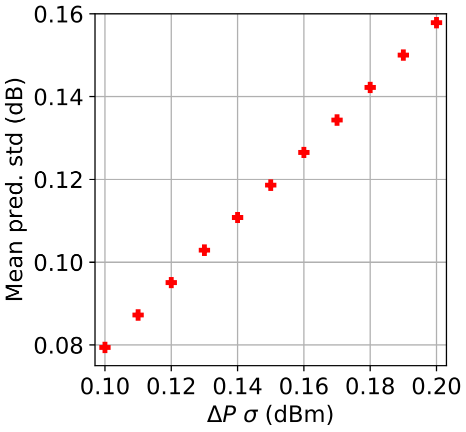


Fig. 9. Variation of the mean of the predictive standard deviation, $\sqrt{V[f_*]}$, of the physics-enhanced GP model with standard deviation of the launch power noise in (6).

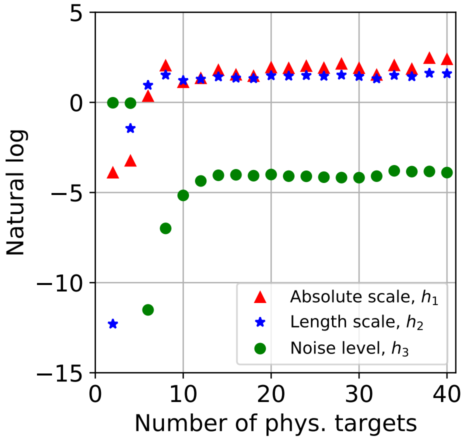


Fig. 10. Variation of optimised kernel hyperparameters with the number of physical model targets used in hyperparameter optimisation. 5 measurements of the SNR have been used in the GP fitting.

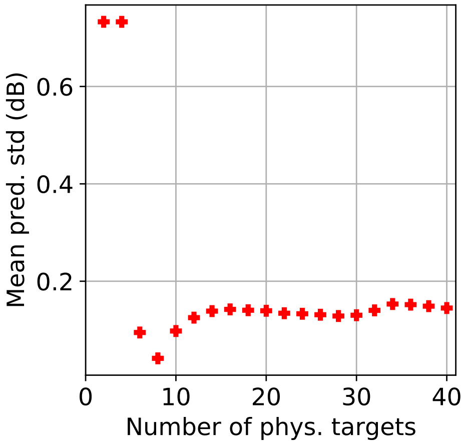


Fig. 11. Variation of the mean of the predictive standard deviation $\sqrt{V[f_*]}$ of the physics-enhanced GP model with the number of physical model targets used in hyperparameter optimisation. 5 measurements of the SNR have been used in the GP fitting.

used in other parts of the model design process. For example, our choice of kernel is motivated by how we expect the signal to behave, based on the functional form of the physical model given in (6). Moreover, the output of the physical model with estimated parameters was used to estimate the launch power corresponding to the optimal SNR, which allowed us to determine a range of launch power over which to measure the SNR. Additionally, we estimate the transceiver back-to-back SNR from [30], thus we have utilised a priori knowledge from the literature.

As outlined in [8], explainability requires some degree of interpretability and transparency, as well as the incorporation of a priori knowledge. An explanation consists of identifying the relevant features that have contributed to a given model prediction [32], which in turn requires model transparency and an interpretation of the model output. We have outlined above why the proposed model is transparent and interpretable, as well as how it incorporates a priori knowledge with machine learning. Thus, we conclude that the proposed method is explainable.

Finally, we emphasise that this method could be applied to regression problems across many domains within optical fiber communications and beyond. All that is required is a prior understanding of the signal under investigation and measured data from the system. In this work, the domain-specific choices are the kernel function used and the physical model of the system, whereas the multi-task learning methodology for incorporating this physical model with the GP is highly general. Thus, to apply physics-informed GPs for a given regression problem, only a suitable kernel and physical model must be specified.

VI. Conclusions

In this work we have outlined a methodology for physics-informed GP regression, creating an explainable approach in which a priori physical models are included in machine learning. Specifically, this is achieved using multi-task learning, in which a physical model is used to generate targets to be used alongside the measured data targets to optimise the hyperparameters. Crucially, this assumes that the physical model targets and data targets can be described by the same underlying statistical distribution, and thus we include uncertainty present in the experimental system under study such that this assumption is valid. As GPs have a well-quantified prediction uncertainty, they are attractive in the context of the high availability requirements of the optical fiber communications domain, where model errors can be catastrophic. We demonstrate that the proposed method facilitates Bayesian inference of the signal variation with fewer measurements of the SNR, allowing us to train GPs in situations in which the number of available measurements is likely to be strongly constrained. Specifically, we show that the physical-model enhanced GP trained using 5 measurements achieves a MAE with respect to the full 13 measured data points only 0.02 dB higher than a standard GP trained on 13 measurements. This increases to a MAE of 0.06 dB with 3 measurements used in fitting, along with a 4.7-fold drop in the model confidence. It is important to note the distinction between the proposed method and using the physical model to generate new data points to use in both hyperparameter optimisation and fitting, as such a method would be very sensitive to errors in the physical model. Furthermore, in this work we consider the explainability of the proposed approach. More specifically, we outline how the physics-informed GPs are transparent and interpretable, and comment on how a priori knowledge is included in the machine learning model itself. Therefore, we conclude that the proposed model is explainable and thus is more likely to be adopted by the telecommunications industry, where optical light paths are established with high availabilities and thus the outputs of machine learning models must be well-understood. Finally, we remark that the proposed method is general and can be applied to regression problems from any domain in which approximate physical models of the target signal are known a priori. This applicability extends beyond the simple one-dimensional input space used in this work, which has been selected in order to demonstrate the physics-informed GP as clearly as possible.

Acknowledgment

This work was supported by the Centre for Doctoral Training in Integrated Photonic and Electronic Systems, funded by the UK Engineering and Physical Sciences Research Council grant EP/L015455/1, and by TRANSNET, funded by the Engineering and Physical Sciences Research Council grant EP/R035342/1. The authors thank Ciena for loan of the equipment used, funding and technical and measurement support. The authors would also like to thank David Saad and Nikita Shevchenko for fruitful discussions relating to this work.

References

[1] F. Musumeci et al. An overview on application of machine learning techniques in optical networks. *IEEE Commun. Surv. Tutor.*, 21:1383–1408, 2018.

[2] F. Mata, I. de Miguel, R. J. Duran, N. Merayo, S. K. Singh, A. Jukan, and M. Chamania. Artificial intelligence (ai) methods in optical networks: A comprehensive survey. *Opt. Switch. Netw.*, 28:43–57, 2018.

[3] P. Poggiolini. The GN model of non-linear propagation in uncompensated coherent optical systems. *J. Light. Technol.*, 30:3857–3879, 2012.

[4] E. Ip and J. M. Kahn. Compensation of dispersion and nonlinear impairments using digital backpropagation. *J. Light. Technol.*, 26:3416–3425, October 2008.

[5] M. Filer et al. Low-margin optical networking at cloud scale. *J. Opt. Commun. Netw.*, 11:C94–C107, September 2019.

[6] Y. Pointurier. Machine learning techniques for quality of transmission estimation in optical networks. *J. Opt. Commun. Netw.*, 13:B60–B71, 2021.

[7] W. Fuhl et al. Explainable online validation of machine learning models for practical applications. In *25th International Conference on Pattern Recognition*, pages 3304–3311, 2021.

[8] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, February 2020.

[9] D. Mackay. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, pages 133–166, 1998.

[10] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT press, Cambridge, MA, 2006.

[11] F. Meng et al. Field trial of gaussian process learning of function-agnostic channel performance under uncertainty. In *Optical Fiber Communication Conference*, pages W4F–5, San Diego, CA, USA, March 2018.

[12] J. Wass, J. Thrane, M. Piels, R. Jones, and D. Zibar. Gaussian process regression for wdm system performance prediction. In *Optical Fiber Communication Conference*, pages 1–3, Los Angeles, CA, USA, March 2017.

[13] T. Panayiotou, S.P. Chatzis, and E. Georgios. Leveraging statistical machine learning to address failure localization in optical networks. *J. Opt. Commun. Netw.*, 2018.

[14] E. Seve et al. Learning process for reducing uncertainties on network parameters and design margins. *J. Opt. Commun. Netw.*, 10:A298–A306, February 2018.

[15] Q. Zhuge et al. Application of machine learning in fiber nonlinearity modeling and monitoring for elastic optical networks. *J. Light. Technol.*, 37:3055–3063, 2019.

[16] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.

[17] X. Jiang et al. Solving the nonlinear schrödinger equation in optical fibers using physics-informed neural network. In *To be published in Proc. Optical Fiber Communications Conference*, 2021.

[18] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, Cambridge, MA, 2018.

[19] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23:550–560, 1997.

[21] P. Virtanen et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[22] T. P. Minka and R. W. Picard. Learning how to learn is learning with point sets. *Unpublished manuscript. Available at https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.7393&rep=rep1&type=pdf*, 1997.

[23] D. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[24] D. J. Ives, H.-M. Chin, F. J. Vaquero Caballero, and S. J. Savory. Single channel probe utilizing the EGN model to estimate link parameters for network abstraction. In *European Conference on Optical Communication*, pages 1–3, Gothenburg, Sweden, September 2017.

[25] D. J. Ives, B. C. Thomsen, R. Maher, and S. J. Savory. Estimating osnr of equalised qpsk signals. *Opt. Express*, 19:B661–B666, 2011.

[26] H.-M. Chin et al. Probabilistic design of optical transmission systems. *J. Light. Technol.*, 35:931–940, 2017.

[27] J. Pesic, N. Rossi, and T. Zami. Impact of margins evolution along ageing in elastic optical networks. *J. Light. Technol.*, 37:4081–4089, 2019.

[28] A. Carena et al. Egn model of non-linear fiber propagation. *Opt. express*, 22:16335–16362, 2014.

[29] P. Poggiolini et al. A simple and effective closed-form gn model correction formula accounting for signal non-gaussian distribution. *J. Light. Technol.*, 33:459–473, 2015.

[30] K. Roberts et al. Performance of dual-polarization QPSK for optical transport systems. *J. Light. Technol.*, 27:3546–3559, August 2009.

[31] Z. C. Lipton. The mythos of model interpretability. *ACM Queue*, 16:31–57, May 2018.

[32] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digit Signal Process*, 73:1–15, February 2018.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin. ”why should i trust you?” explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[34] S. Lundberg and S-I. Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[35] L. von Rueden et al. Informed machine learning–A Taxonomy and Survey of Integrating Knowledge into Learning Systems. *arXiv preprint arXiv:1903.12394*, 2019.