Hierarchical Bayesian Approach to Experimental Data Fusion: Application to Strength Prediction of High Entropy Alloys from Hardness Measurements

Sharmila Karumuri[1], Zachary D. McClure[2], Alejandro Strachan[2], Michael Titus[2], and Ilias Bilionis*[1]

[1]School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, USA
[2]School of Materials Engineering, Purdue University, West Lafayette, IN 47907, USA

October 26, 2023

Abstract

The discovery of materials with improved properties can be accelerated by models with the ability to combine data from multiple experimental information sources. A recurring task in the toolbox of practitioners is to map input physical descriptors to output properties of interest. Typically, both the outputs and many of the inputs are experimentally measured and, thus, noisy. Probabilistic regression methods, e.g., Gaussian process regression, can easily deal with noisy outputs, even if the noise is input-dependent. However, most regression methods cannot process noisy inputs. Ignoring input uncertainty leads to inaccurate predictive uncertainty, a crucial ingredient for the sequential design of experiments. The objective of this paper is to develop a regression methodology that can deal with input uncertainty when one wishes to correlate an inexpensive experimental measurement (e.g., hardness) to an expensive one (e.g., yield strength). Our hierarchical Bayesian approach uses two Gaussian processes. The first one maps noiseless physical descriptors to the inexpensive experimental measurement. The second Gaussian process maps noiseless physical descriptors and the inexpensive experimental measurement to the expensive experimental measurement. The two Gaussian processes form a nested model that is not analytically tractable. To overcome this issue, we propose semi-analytical approximations to both the marginal likelihood and the posterior predictive distribution. The result is a model that is practical to train and use. We demonstrate the merits of the proposed method through a synthetic dataset in which we control all the uncertainties. The statistical tests clearly show that standard Gaussian process regression cannot cope with input uncertainty whereas our proposed method consistently yields better predictive distributions. Finally, we apply the method to the task of predicting the yield strength of high entropy alloys from hardness on an exhaustive dataset compiled from the available literature.

Keywords: Noisy inputs; Input uncertainty; Hierarchical Gaussian process regression; High entropy alloys; Yield strength prediction; Hardness.

*Corresponding author: ibilion@purdue.edu

1 Introduction

High entropy alloys (HEAs) formed by mixing relatively equal proportions of four or more elements are drawing the interests of researchers after their introduction in 2004 [1, 2]. These alloys have promising oxidation resistance, strength performance at high temperatures and ductility at low temperatures making them promising candidates for structural material applications [3, 4, 5, 6, 7]. Predicting the properties of these alloys using experimental techniques is costly and time consuming. Therefore, data-driven techniques such as machine learning (ML) are being utilized to predict the mechanical properties of new HEAs [8, 9, 10, 11].

Typically, some of the inputs and outputs used in building ML models for HEAs are experimental quantities and, thus, noisy. For example, in the strength prediction of HEA alloys using hardness as an input, both the output strength data and input hardness data are experimentally measured quantities. Data uncertainty is not taken into account by common regression techniques like kernel ridge regression [12] and neural networks [13]. Whereas statistical regression techniques like Bayesian linear regression [14, 15], Gaussian process regression (GPR)[16, 15] and Bayesian neural networks (BNNs) [17, 18, 19], do account for output uncertainty even when the noise is input-dependent (heteroscedasticity [20, 21, 22]). However, these methods rely on the assumption that the inputs are noise-free, which is not valid in many material science applications. Ignoring this input uncertainty may affect the quality of the model, resulting in an inaccurate predictive distribution. It is important to capture predictive mean and uncertainty estimates accurately as they are key ingredients for the sequential design of experiments using methods like Bayesian global optimization [23]. Hence, the objective of this paper is to develop a regression methodology that is capable of dealing with input uncertainty when one wishes to correlate an inexpensive experimental measurement (hardness) to an expensive one (yield strength). In [11], authors correlated oxidation stiffness available via *ab initio* simulations to expensive melting temperature using random-forest based surrogate models, however, these models account for uncertainty in a non-Bayesian way.

There are several examples of regression methods that can deal with input uncertainty. The authors of [24] employ a modified least-squares method in which they modify the loss of the regression problem to account for errors due to noise on the outputs as well as the inputs. Several researchers have put forth various ways of building GPR models from uncertain inputs. Specifically, in [25] the authors proposed a modified Gaussian process model in which the covariance function has been corrected to account for noise variance in the input. In [26] uncertain inputs are assumed to follow a Gaussian distribution with known variance and a covariance matrix is constructed by marginalizing over the uncertain inputs. The authors of [27] proposed a noisy input GPR method in which they treat the input data as though they were deterministic and they amplify the corresponding output variance to account for the input noise. On the neural network front, the authors in [28] proposed an extension of BNN to deal with noisy inputs, but the drawback of this approach is that the estimation of the predictive distribution requires sampling by Markov chain Monte Carlo (MCMC) methods [15, 29] which is computationally intensive.

The majority of the approaches mentioned above are based on the idea of integrating out the uncertain inputs and coming up with different approximations of intractable integrals. This is absolutely necessary when the inputs are independent and noisy. However, in the typical materials application of interest here, we encounter a slightly different situation. Namely, we have some available noise-free physical descriptors that correlate with the noisy input of interest. This allows us to develop an approach similar to [30], in which one explicitly builds a regression model that connects the noise-free physical descriptors to the noisy input. This approach has two advantages. First, it is much more tractable than integrating over all input uncertainty. Second, it produces as a byproduct a de-noised version of the noisy inputs, which may be of

interest independently.

We differ from [30] in the sense that we follow a hierarchical Bayesian approach [31]. In particular, we use two Gaussian processes: (i) one that connects the noiseless physical descriptors to the inexpensive experimental measurement, and (ii) a second one that maps the noiseless physical descriptors and the inexpensive experimental measurement to the expensive experimental measurement. To deal with this nested and analytically intractable model, we derive semi-analytical approximations of the marginal likelihood and the posterior predictive distribution. We call our method hierarchical Gaussian process regression (HGPR). We compare HGPR with standard GPR (SGPR) on synthetic examples in which we control all the uncertainties. Finally, we apply HGPR to the problem of predicting the yield strength of high entropy alloys from hardness data using a dataset compiled from the available literature.

The rest of the paper is structured as follows. In Sec. 2 we present our methodology. We start in Sec. 2.1 by defining our problem mathematically. Then, we discuss the SGPR method in Sec. 2.2 and our proposed method in Sec. 2.3. In Sec. 2.4, we describe the diagnostics used for the model validation. We present our results in Sec. 3. Specifically, in Sec. 3.1 we compare methods SGPR with HGPR on synthetic data and in Sec. 3.2, we illustrate the effectiveness of our method in yield strength prediction of high entropy alloys from hardness measurements. We present our concluding remarks in Sec. 4.

2 Methodology

2.1 Problem definition

We are interested in predicting a physical quantity of interest $y \in \mathbb{R}$ that is available through relatively expensive experiments. Each experiment is described by a vector of noise-free physical descriptors $x \in \mathbb{R}^d$. In our running example of high-entropy alloys, the physical quantity of interest y is the yield strength and the physical descriptors x include 25 quantities estimated from the properties of elements in the alloy and the phase information of the alloy (see Appendix A.1). Unfortunately, one can have only a handful of y measurements as they are expensive. Therefore, since x is high-dimensional, it is impossible to learn the map from physical descriptors to the physical quantity of interest.

Now consider the following scenario in which we have access to inexpensive experimental measurements of another physical quantity, $z \in \mathbb{R}$, which correlates with y. In our running example, the physical quantity z is the hardness. The typical approach is to learn the map f from x and z to y via the SGPR method using the data $\mathcal{D}_f = \Big((\mathbf{X}_f, \mathbf{z}_f), \mathbf{y}_f\Big)$, where $\mathbf{X}_f = (x_{f,1}, \ldots, x_{f,N_f})$, $\mathbf{z}_f = (z_{f,1}, \ldots, z_{f,N_f})$, and $\mathbf{y}_f = (y_{f,1}, \ldots, y_{f,N_f})$ are N_f observations of all relevant quantities. Notice that here both the inputs $\mathbf{z}_f$ and the outputs $\mathbf{y}_f$ are noisy as they are experimentally measured, but SGPR assumes that only the outputs are noisy. Disregarding the uncertainty in $\mathbf{z}_f$ leads to inaccurate predictive distributions. So we developed a regression method HGPR that solves this problem by accounting for all the uncertainties in an hierarchical manner by first conducting enough inexpensive z experiments to learn the map g from x to z using the data $\mathcal{D}_g = (\mathbf{X}_g, \mathbf{z}_g)$, where $\mathbf{X}_g = (x_{g,1}, \ldots, x_{g,N_g})$, and $\mathbf{z}_g = (z_{g,1}, \ldots, z_{g,N_g})$, and then use this learned de-noised input for learning the required map f. Note that $\mathcal{D}_g$ should include all observations of z in $\mathcal{D}_f$.

From now on we denote all the data from expensive experiments and inexpensive experiments as $\mathcal{D} = (\mathcal{D}_g, \mathcal{D}_f)$. Next, in Sec. 2.2 and Sec. 2.3 we describe in detail how the required map f is learnt using both SGPR and our HGPR methods, followed by the overall Algorithm 1 depicting the model building process using the HGPR method.

2.2 Standard Gaussian process regression

Given a dataset $\mathcal{D}_f$, we can learn the map f using SGPR method as shown in many previous works [32, 33, 34, 35]. In SGPR (see Fig. 1a), the input data $\mathbf{z}_f$ in $\mathcal{D}_f$ is assumed to be noiseless and the noise in the output data $\mathbf{y}_f$ is approximated by Gaussian noise with constant variance γ_y^2:

$$\mathbf{y}_f|f, \mathbf{X}_f, \mathbf{z}_f, \gamma_y \sim \mathcal{N}(f(\mathbf{X}_f, \mathbf{z}_f), \gamma_y^2\mathbf{I}).$$

One puts a prior over the space of functions f:

$$f \sim \mathrm{GP}(0, k_f),$$

which encodes the beliefs about regularity and length-scales of f. These beliefs are encoded using a mean function typically chosen to be zero and a covariance function k_f with parameters ϕ_f. Then, one finds the parameters ϕ_f and γ_y by maximizing the marginal likelihood of the data $p(\mathbf{y}_f|\mathbf{X}_f, \mathbf{z}_f, \phi_f, \gamma_y)$, which is Gaussian and given by:

$$p(\mathbf{y}_f|\mathbf{X}_f, \mathbf{z}_f, \phi_f, \gamma_y) = \mathcal{N}(\mathbf{y}_f|0, \mathbf{K}_f + \gamma_y^2\mathbf{I}),$$

where $\mathbf{K}_f$ is the covariance matrix constructed using the covariance function k_f between every pair of inputs in $\mathcal{D}_f$.

Having identified the model parameters ϕ_f, one conditions the prior measure on the available data. This way, one obtains a posterior probability measure over the space of f's, which is another Gaussian process [16]. This posterior Gaussian process can be used to derive the (point) *posterior predictive distribution* which predicts the output on an arbitrary inputs x_* and z_*. The posterior predictive distribution is a univariate Gaussian and comes in two versions. The first version, which includes only the epistemic uncertainty, predicts the noiseless output f_*:

$$f_*|\mathcal{D}_f, x_*, z_* \sim \mathcal{N}(\tilde{\mu}_f(x_*, z_*), \tilde{\sigma}_f^2(x_*, z_*)), \tag{1}$$

$$\text{where, } \tilde{\mu}_f(x_*, z_*) = \mathbf{k}_f((x_*, z_*), (\mathbf{X}_f, \mathbf{z}_f))^T(\mathbf{K}_f + \gamma_y^2\mathbf{I})^{-1}\mathbf{y}_f,$$
$$\tilde{\sigma}_f^2(x_*, z_*) = k_f((x_*, z_*), (x_*, z_*)) - \mathbf{k}_f((x_*, z_*), (\mathbf{X}_f, \mathbf{z}_f))^T(\mathbf{K}_f + \gamma_y^2\mathbf{I})^{-1}\mathbf{k}_f((x_*, z_*), (\mathbf{X}_f, \mathbf{z}_f)),$$

and $\mathbf{k}_f((x_*, z_*), (\mathbf{X}_f, \mathbf{z}_f)) = [k_f((x_*, z_*), (x_{f,1}, z_{f,1})) \dots k_f((x_*, z_*), (x_{f,N_f}, z_{f,N_f}))]^T$ is the cross covariance vector between the arbitrary inputs and the inputs in $\mathcal{D}_f$. The second version, which includes both epistemic and aleatory uncertainty, predicts the measured output y_*:

$$y_*|\mathcal{D}_f, x_*, z_* \sim \mathcal{N}(\tilde{\mu}_f(x_*, z_*), \tilde{\sigma}_f^2(x_*, z_*) + \gamma_y^2). \tag{2}$$

2.3 Hierarchical Gaussian processes regression

The SGPR method described above completely disregards the measurement uncertainty in the input data $\mathbf{z}_f$ by assuming it to be noiseless. The $\mathbf{z}_f$ data, collected from experiments could be very noisy and since it is substantially correlated with the requisite output data $\mathbf{y}_f$, the resulting SGPR predictions will have inaccurate mean and uncertainty estimates.

To circumvent this problem, in our HGPR method we account for all the uncertainties in the dataset $\mathcal{D}_f$ in a hierarchical manner (see Fig. 1b). Firstly, we learn a map g from x to z using the data $\mathcal{D}_g$ so that we can de-noise the input data $\mathbf{z}_f$. For learning this map g, the noise in the data $\mathbf{z}_g$ is approximated by

Gaussian noise with constant variance γ_z^2:

$$\mathbf{z}_g|g, \mathbf{X}_g, \gamma_z \sim \mathcal{N}(g(\mathbf{X}_g), \gamma_z^2\mathbf{I}).$$

We begin with placing a Gaussian process prior over the space of functions g:

$$g \sim \text{GP}(0, k_g),$$

with a covariance function k_g with parameters ϕ_g encodes the beliefs about regularity and length-scales of the inexpensive physical quantity. Then, we find the parameters ϕ_g and γ_z by maximizing the marginal likelihood of the data,

$$p(\mathbf{z}_g|\mathbf{X}_g, \phi_g, \gamma_z) = \mathcal{N}(\mathbf{z}_g|0, \mathbf{K}_g + \gamma_z^2\mathbf{I}), \tag{3}$$

where $\mathbf{K}_g$ is the covariance matrix constructed using the covariance function k_g between every pair of inputs in $\mathcal{D}_g$.

Having identified the model parameters ϕ_g, we can get the the analytically available Gaussian posterior probability measure [16] for g. Using this posterior Gaussian process, we estimate the two versions of the (point) *posterior predictive distribution* of inexpensive quantity at arbitrary inputs x_*, the version with only epistemic uncertainty, and the other version, which includes both epistemic and aleatoric uncertainty, as follows:

$$\begin{aligned} g_*|\mathcal{D}_g, x_* &\sim \mathcal{N}(\tilde{\mu}_g(x_*), \tilde{\sigma}_g^2(x_*)), \\ z_*|\mathcal{D}_g, x_* &\sim \mathcal{N}(\tilde{\mu}_g(x_*), \tilde{\sigma}_g^2(x_*) + \gamma_z^2), \end{aligned} \tag{4}$$

$$\begin{aligned} \text{where, } \tilde{\mu}_g(x_*) &= \mathbf{k}_g(x_*, \mathbf{X}_g)^T(\mathbf{K}_g + \gamma_z^2\mathbf{I})^{-1}\mathbf{y}_g, \\ \tilde{\sigma}_g^2(x_*) &= k_g(x_*, x_*) - \mathbf{k}_g(x_*, \mathbf{X}_g)^T(\mathbf{K}_g + \gamma_z^2\mathbf{I})^{-1}\mathbf{k}_g(x_*, \mathbf{X}_g), \end{aligned}$$

and $\mathbf{k}_g(x_*, \mathbf{X}_g) = [k_g(x_*, x_{g,1}) \cdots k_g(x_*, x_{g,N_g})]^T$ is the cross covariance vector between the arbitrary input and the inputs in $\mathcal{D}_g$.

Also, note that while we use noiseless descriptors x to learn the de-noised map g of the noisy input, however in principle, we could also use a completely different set of inputs to learn this map (e.g., a subset of descriptors in x to learn g) depending on what z depends on.

Following the learning of this map g, we learn a second map f from x and $g(x)$ to y, using the data $\mathcal{D}$. The noise in the data $\mathbf{y}_f$ is approximated by Gaussian noise with constant variance γ_y^2:

$$\mathbf{y}_f|f, g, \mathbf{X}_f, \gamma_y \sim \mathcal{N}(f(\mathbf{X}_f, g(\mathbf{X}_f)), \gamma_y^2\mathbf{I}).$$

We place a Gaussian process prior over the space of functions f:

$$f|g \sim \text{GP}(0, k_f),$$

with zero mean function and a covariance function k_f defined on the inputs x and $g(x)$. k_f encodes the beliefs about regularity and length-scales of the expensive physical quantity and has parameters ϕ_f. The Gaussian process f (i.e., $f(x, g(x))$) is a deep GP [36], hence the marginal likelihood,

$$p(\mathbf{y}_f|\mathbf{X}_f, \mathcal{D}_g, \phi_f, \gamma_y) = \int p(\mathbf{y}_f|\mathbf{X}_f, g(\mathbf{X}_f), \phi_f, \gamma_y)p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f), \tag{5}$$

and the posterior predictive distribution at arbitrary input x_*,

$$p(f_*|\mathcal{D}, x_*) = \iint p(f_*|\mathbf{y}_f, \mathbf{X}_f, g(\mathbf{X}_f), x_*, g(x_*))p(g(\mathbf{X}_f), g(x_*)|\mathcal{D}_g, x_*)dg(\mathbf{X}_f)dg(x_*), \tag{6}$$

are no longer Gaussian and analytically available, posing significant computational challenges. To overcome these computational challenges, we derive a semi-analytical approximations.

Think of the term $p(\mathbf{y}_f|\mathbf{X}_f, g(\mathbf{X}_f), \phi_f, \gamma_y)$ in Eq. (5) as a function $h(g(\mathbf{X}_f))$. Then Eq. (5) reduces to,

$$p(\mathbf{y}_f|\mathbf{X}_f, \mathcal{D}_g, \phi_f, \gamma_y) = \int h(g(\mathbf{X}_f))p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f). \tag{7}$$

Expand $h(g(\mathbf{X}_f))$ in a first-order Taylor series around the predictive mean $\tilde{\mu}_g(\mathbf{X}_f)$ from Eq. (4) gives,

$$h(g(\mathbf{X}_f)) \approx h(\tilde{\mu}_g(\mathbf{X}_f)) + \nabla h(\tilde{\mu}_g(\mathbf{X}_f)) \left[g(\mathbf{X}_f) - \tilde{\mu}_g(\mathbf{X}_f)\right].$$

Substituting this expansion in Eq. (7) gives,

$$\begin{aligned}
p(\mathbf{y}_f|\mathbf{X}_f, \mathcal{D}_g, \phi_f, \gamma_y) &\approx \int \Big[h(\tilde{\mu}_g(\mathbf{X}_f)) + \nabla h(\tilde{\mu}_g(\mathbf{X}_f))(g(\mathbf{X}_f) - \tilde{\mu}_g(\mathbf{X}_f))\Big]p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f) \\
&= h(\tilde{\mu}_g(\mathbf{X}_f)) \int p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f) + \nabla h(\tilde{\mu}_g(\mathbf{X}_f)) \int (g(\mathbf{X}_f) - \tilde{\mu}_g(\mathbf{X}_f))p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f) \\
&= h(\tilde{\mu}_g(\mathbf{X}_f)) \cdot 1 + \nabla h(\tilde{\mu}_g(\mathbf{X}_f))(\tilde{\mu}_g(\mathbf{X}_f) - \tilde{\mu}_g(\mathbf{X}_f) \cdot 1) \\
&= h(\tilde{\mu}_g(\mathbf{X}_f)).
\end{aligned}$$

Therefore:

$$p(\mathbf{y}_f|\mathbf{X}_f, \mathcal{D}_g, \phi_f, \gamma_y) \approx p(\mathbf{y}_f|\mathbf{X}_f, g(\mathbf{X}_f) = \tilde{\mu}_g(\mathbf{X}_f), \phi_f, \gamma_y) = \mathcal{N}(\mathbf{y}_f|0, \tilde{\mathbf{K}}_f + \gamma_y^2\mathbf{I}) \tag{8}$$

where $\tilde{\mathbf{K}}_f$ is the covariance matrix defined between every pair of inputs in $(\mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f))$ using the covariance function k_f.

We approximate the required posterior predictive distribution with epistemic uncertainty at arbitrary input x_* using Monte Carlo method [37, 38] by iteratively sampling from the posterior predictive distribution of g_* as follows:

$$\begin{aligned}
p(f_*|\mathcal{D}, x_*) &= \iint p(f_*|\mathbf{y}_f, \mathbf{X}_f, g(\mathbf{X}_f), x_*, g(x_*))p(g(\mathbf{X}_f), g(x_*)|\mathcal{D}_g, x_*)dg(\mathbf{X}_f)dg(x_*) \\
&\approx \iint p(f_*|\mathbf{y}_f, \mathbf{X}_f, g(\mathbf{X}_f), x_*, g(x_*))p(g(\mathbf{X}_f)|\mathcal{D}_g)p(g(x_*)|\mathcal{D}_g, x_*)dg(\mathbf{X}_f)dg(x_*) \\
&= \int \Big[\int p(f_*|\mathbf{y}_f, \mathbf{X}_f, g(\mathbf{X}_f), x_*, g(x_*))p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f)\Big]p(g(x_*)|\mathcal{D}_g, x_*)dg(x_*)
\end{aligned} \tag{9}$$

Now let the term in the square parentheses be 'L' and $p(f_*|\mathbf{y}_f, \mathbf{X}_f, g(\mathbf{X}_f), x_*, g(x_*))$ be a function $w(g(\mathbf{X}_f))$. Then 'L' reduces to,

$$L = \int w(g(\mathbf{X}_f))p(g(\mathbf{X}_f)|\mathcal{D}_g)dg(\mathbf{X}_f),$$

Similar to before, expanding $w(g(\mathbf{X}_f))$ in first-order Taylor series around predictive mean $\tilde{\mu}_g(\mathbf{X}_f)$ from Eq. (4) would give,

$$L \approx w(\tilde{\mu}_g(\mathbf{X}_f)) = p(f|\mathbf{y}_f, \mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f), x_*, g(x_*)).$$

Substituting this 'L' in Eq. (9) gives,

$$\begin{aligned} p(f_*|\mathcal{D}, x_*) &\approx \int p(f_*|\mathbf{y}_f, \mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f), x_*, g(x_*)) p(g(x_*)|\mathcal{D}_g, x_*) dg(x_*) \\ &= \int p(f_*|\mathbf{y}_f, \mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f), x_*, g_*) p(g_*|\mathcal{D}_g, x_*) dg_* \\ &\approx \frac{1}{N} \sum_{i=1}^{N} p(f|\mathbf{y}_f, \mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f), x_*, g_*^{(i)}), \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(\tilde{\mu}_f(x_*, g_*^{(i)}), \tilde{\sigma}_f^2(x_*, g_*^{(i)})), \end{aligned} \tag{10}$$

where, $g_*^{(i)}$ are independent samples from the predictive distribution of inexpensive quantity in Eq. (4) and

$$\begin{aligned} \tilde{\mu}_f(x_*, g_*^{(i)}) &= \mathbf{k}_f \left((x_*, g_*^{(i)}), (\mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f)) \right)^T (\tilde{\mathbf{K}}_f + \gamma_y^2 \mathbf{I})^{-1} \mathbf{y}_f, \\ \tilde{\sigma}_f^2(x_*, g_*^{(i)}) &= k_f((x_*, g_*^{(i)}), (x_*, g_*^{(i)})) \\ &\quad - \mathbf{k}_f \left((x_*, g_*^{(i)}), (\mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f)) \right)^T (\tilde{\mathbf{K}}_f + \gamma_y^2 \mathbf{I})^{-1} \mathbf{k}_f \left((x_*, g_*^{(i)}), (\mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f)) \right), \end{aligned}$$

where,

$$\mathbf{k}_f \left((x_*, g_*^{(i)}), (\mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f)) \right) = [k_f((x_*, g_*^{(i)}), (x_{f,1}, \tilde{\mu}_g(x_{f,1}))) \cdots k_f((x_*, g_*^{(i)}), (x_{f,N_f}, \tilde{\mu}_g(x_{f,N_f})))]^T.$$

Following similar approach, the (point) *posterior predictive distribution* of expensive quantity with epistemic and aleatory uncertainty reduces to,

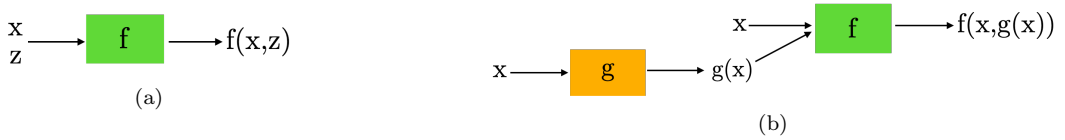
$$\begin{aligned} p(y_*|\mathcal{D}, x_*) &\approx \frac{1}{N} \sum_{i=1}^{N} p(y_*|\mathbf{y}_f, \mathbf{X}_f, \tilde{\mu}_g(\mathbf{X}_f), x_*, g_*^{(i)}), \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(\tilde{\mu}_f(x_*, g_*^{(i)}), \tilde{\sigma}_f^2(x_*, g_*^{(i)}) + \gamma_y^2), \end{aligned} \tag{11}$$

where $g_*^{(i)} \sim g_*|\mathcal{D}_g, x_*$ (see Eq. (4)).

Algorithm 1 Hierarchical Gaussian process regression (HGPR)

Require: Training datasets $\mathcal{D}_g = (\mathbf{X}_g, \mathbf{z}_g)$ and $\mathcal{D}_f = ((\mathbf{X}_f, \mathbf{z}_f), \mathbf{y}_f)$

1: To learn the de-noised response of the noisy input, train the Gaussian process g with $\mathcal{D}_g$ by maximizing the marginal likelihood in Eq. (3).
2: Estimate the mean of the response map g at $\mathbf{X}_f$ as $\tilde{\mu}_g(\mathbf{X}_f)$ using Eq. (4).
3: Train the Gaussian process f with $\mathcal{D} = (\mathcal{D}_g, \mathcal{D}_f)$ by maximizing the marginal likelihood in Eq. (8) using $\tilde{\mu}_g(\mathbf{X}_f)$. Estimate the two versions of posterior predictive distribution as in Eq. (10) and Eq. (11) respectively.



(a) (b)

Figure 1: 1a corresponds to the schematic view of the model in SGPR method, where the response surface f is trained on the dataset $\mathcal{D}_f$. 1b corresponds to the schematic view of the model in HGPR method, where the response surface g and f are trained on the datasets $\mathcal{D}_g$ and $\mathcal{D}$ respectively.

2.4 Evaluation metrics

Let $\mathcal{D}_v = ((\mathbf{X}_v, \mathbf{z}_v), \mathbf{y}_v)$, be the validation data used to compare the models learned from the two methods, SGPR and HGPR, where $\mathbf{X}_v = (x_{v,1}, \ldots, x_{v,N_v})$, $\mathbf{z}_v = (z_{v,1}, \ldots, z_{v,N_v})$, and $\mathbf{y}_v = (y_{v,1}, \ldots, y_{v,N_v})$ are the N_v observations of all the relevant quantities.

We evaluate the mean absolute error between the validation data and the mean of the predictive distribution as follows:

$$\text{MAE} = \frac{1}{N_v} \sum_{i=1}^{N_v} \left| y_{v,i} - \mathbb{E}\Big[f_* | \mathcal{D}, x_* = x_{v,i}, z_* = z_{v,i} \Big] \right|, \tag{12}$$

but note that this metric is not reliable when the validation data are very noisy. In instances where noiseless validation data is available, we evaluate the absolute error in Eq. (12) with respect to the noiseless validation data and we refer to this metric as $\text{MAE}_{\text{truth}}$.

Both MAE and $\text{MAE}_{\text{truth}}$ are only point-based metrics and they do not validate the complete predictive distribution of the models. Being able to capture the predictive distribution correctly is important for designing active learning schemes that can explore the input space. To validate the full predictive distribution of the models, we employ statistical tests. In SGPR method, we do this by checking whether standardized errors (as in [39]) follow a standard normal distribution. But we cannot utilize this statistical test here because the predictive distribution from the HGPR method is no longer Gaussian.

We developed a different statistical test based on probability integral transform principle (see [40]). The principle states that if T is a random variable with cumulative distribution function (CDF) $F_T(T)$, then the random variable R defined as $R = F_T(T)$ has a uniform distribution. Based on this idea, we validate the model by checking whether the CDF of our predictive model evaluated at the validation data follows the uniform distribution. That is, we test if

$$k_i = F\Big[y_* = y_{v,i} | \mathcal{D}, x_* = x_{v,i}, z_* = z_{v,i} \Big] \sim U(0,1), \tag{13}$$

where, $F[y_* | \mathcal{D}, x_*, z_*]$ is the empirical cumulative distribution function (ECDF) of the predictive model estimated from the samples of the predictive distribution. In other words, Eq. (13) lets us test whether validation data is arising from the predictive distribution given by our model.

We check this diagnostic in two ways - 1) Using Kolmogorov–Smirnov test statistic (KS test statistic) [41, 42]; This statistic quantifies the distance between the ECDF of the sample k_i's and the CDF of the uniform distribution. When $k_{1:N_v}$ follows a perfect uniform distribution then the KS test statistic should be zero. 2) Using the quantile-quantile plot (Q-Q plot) [43]; Here we compare the empirical quantiles of the k_i's to those of the uniform quantiles. When $k_{1:N_v}$ follows uniform distribution, the q-q plot falls on the 45° line that crosses the axes.

3 Results

In this section, we validate our approach using a synthetic example and a realistic material science application. For all the cases, we present the results from employing our HGPR method as well as SGPR method for comparison and validation.

The examples considered here were chosen in order to highlight the benefits of using our hierarchical approach and its effectiveness in building a model by fusing information from different sources along with accounting for the input uncertainty as well as output uncertainty. Throughout all the examples we use the squared exponential kernel function [16] with automatic relevance determination (ARD) of weights [16, 15]. ARD of weights corresponds to different length-scales for each dimension in the input and thereby letting method to detect which input variables have more effect on the predictive distribution. All models are implemented using the open source library GPy [44].

3.1 Example 1: Pedagogical example

Consider the case where the true responses governing the inexpensive and expensive physical quantities, g and f, respectively, follow these equations:

$$\begin{aligned} g(x) &= \sin(8\pi x), \\ f(x) &= f(x, g(x)) = (x - \sqrt{2})g(x)^2, \end{aligned} \tag{14}$$

where response g is chosen to be a sinusoidal wave with four periods and response f is obtained by transforming g non-linearly. See Fig. 2 for an illustration. Assume that we have access only to a finite number of noisy observations of g and f. In particular, we obtain the observations of z and y in $\mathcal{D}_f$ by randomly sampling the true responses g and f with additive Gaussian noise of variances $\gamma_z^2 = 0.25^2$ and $\gamma_y^2 = 0.2^2$, respectively at $N_f = 25$ points. Similarly, we generate dataset $\mathcal{D}_g$ of inexpensive observations by randomly sampling the true responses g with additive Gaussian noise of variance γ_z^2 at $N_g = 65$ points. Note that, while sampling we ensure that $\mathcal{D}_g$ includes all observations of z in $\mathcal{D}_f$. Now, given the availability of these datasets $\mathcal{D}_g$ and $\mathcal{D}_f$, the goal is to learn the true response f governing the expensive physical quantity as accurately as possible. Also, since we have access to the true responses here, we generate noiseless validation data of $N_v = 150$ points by randomly sampling in the domain $[0, 1]$ to compare the results from the two approaches.

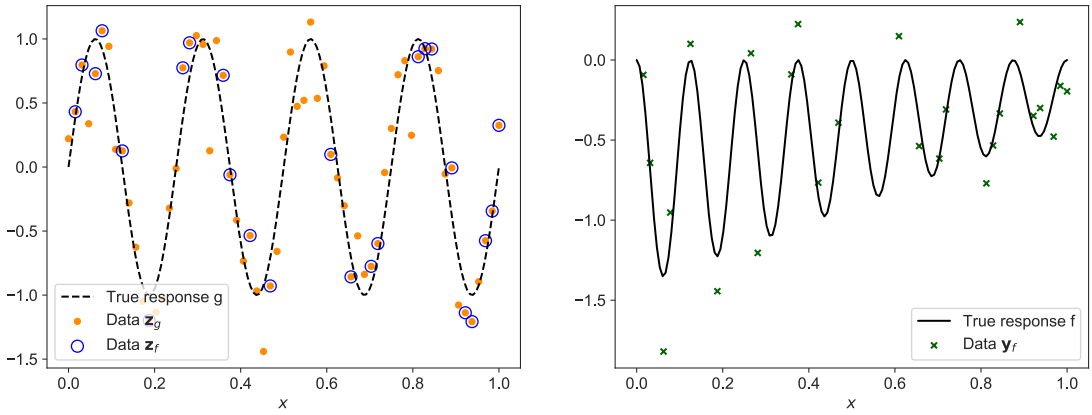


Figure 2: Example 1 - Left and the right plots corresponds to noisy data of inexpensive and expensive physical quantities respectively obtained by randomly sampling.

The first approach is to learn the required map f following the SGPR method discussed in Sec. (2.2) using dataset $\mathcal{D}_f$ (but note that this method assumes input data $\mathbf{z}_f$ in $\mathcal{D}_f$ to be noiseless). We assumed that the noise variance γ_y is known. As seen in Fig. 3, this approach does not provide a reasonable reconstruction of f; the mean response is inaccurate and noisy, and also the uncertainty is overestimated. Note that, SGPR method does not utilize the extra inexpensive measurements available in $\mathbf{z}_g$ for learning the expensive quantity response.

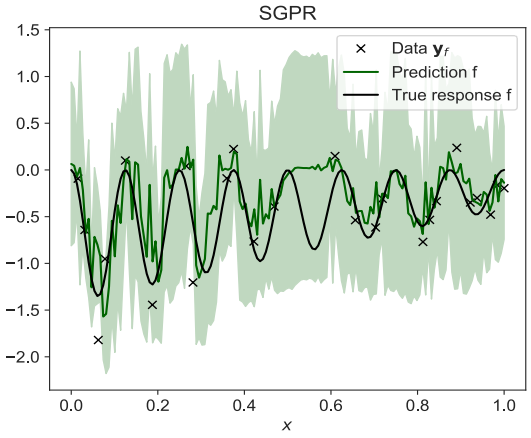
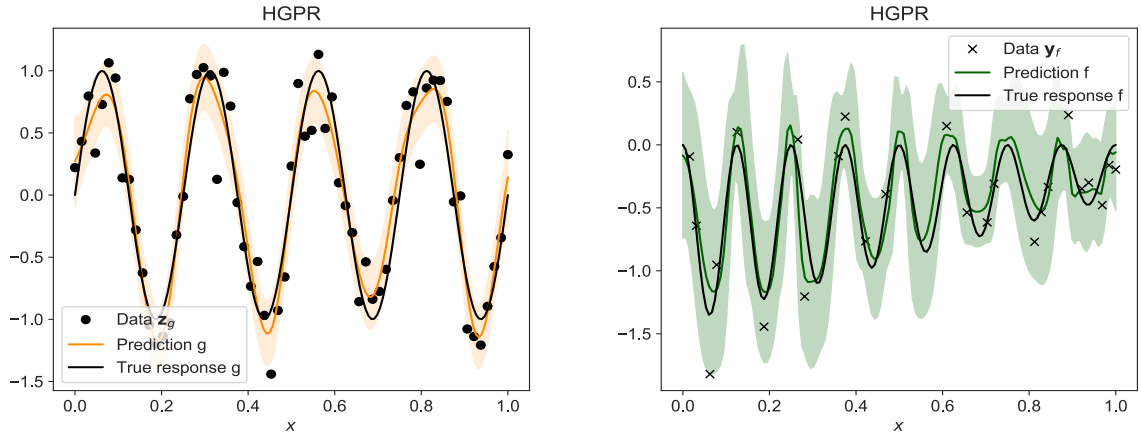


Figure 3: Example 1 - Predictive distribution from the SGPR method. Green line corresponds to mean response and shaded region corresponds to 1.96 standard deviation band of f ($\text{MAE}_{\text{truth}} = 0.25$, KS test statistic $= 0.26$).

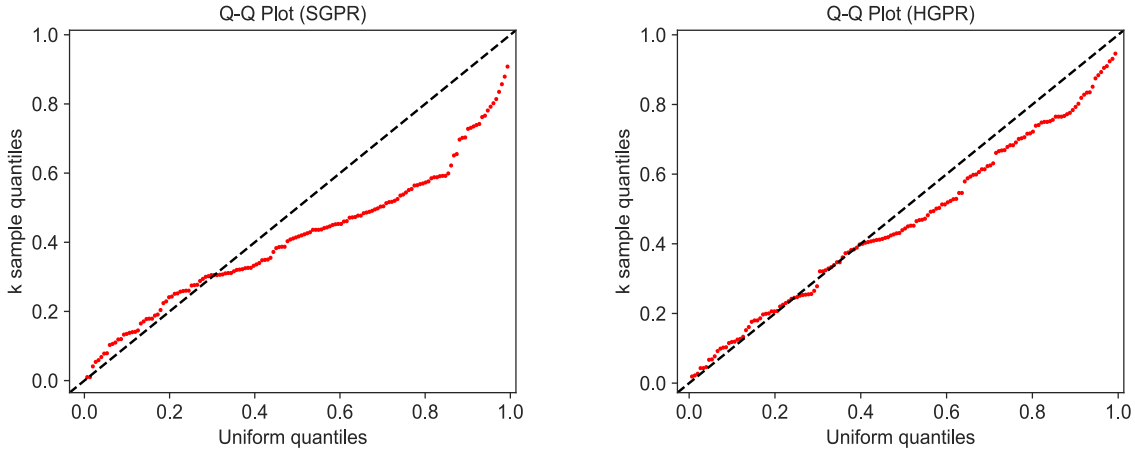
Next, we present the result using our proposed HGPR method discussed in Sec. (2.3). We begin by first learning a response map g using $\mathcal{D}_g$. Then we use this learned de-noised input data $g(\mathbf{X}_f)$ in the place of $\mathbf{z}_f$ for learning the required map f. It is assumed that the noise variances γ_z and γ_y are known. As seen in Fig. 4, this approach provides an accurate reconstruction of f, i.e., mean response and uncertainty estimates are sensible. This can be evidently seen from the low values of our validation metrics (see Sec. (2.4)), i.e., the $\text{MAE}_{\text{truth}}$ and KS test statistic using HGPR method compared to the SGPR method in Tab. 1. Also,

285 from Fig. 5b it is evident that Q-Q plot of $k_{1:N_v}$ from the HGPR method follows a 45° line compared to the Q-Q plot of $k_{1:N_v}$ from the SGPR (Fig. 5a) method validating the predictive distribution.



(a) (b)

Figure 4: Example 1 - Predictive distributions from the HGPR method. Left plot orange line corresponds to mean response and shaded region corresponds to 1.96 standard deviation band of g and in the right plot green line corresponds to mean response and the shaded region corresponds to the 1.96 standard deviation band of f ($\text{MAE}_{\text{truth}} = 0.12$, KS test statistic $= 0.11$).



(a) (b)

Figure 5: Example 1 - Left plot and right plot corresponds to Q-Q plot of $k_{1:N_v}$ using SGPR and HGPR methods respectively.

Table 1: Example 1 - Comparison of the validation metrics from the two methods.

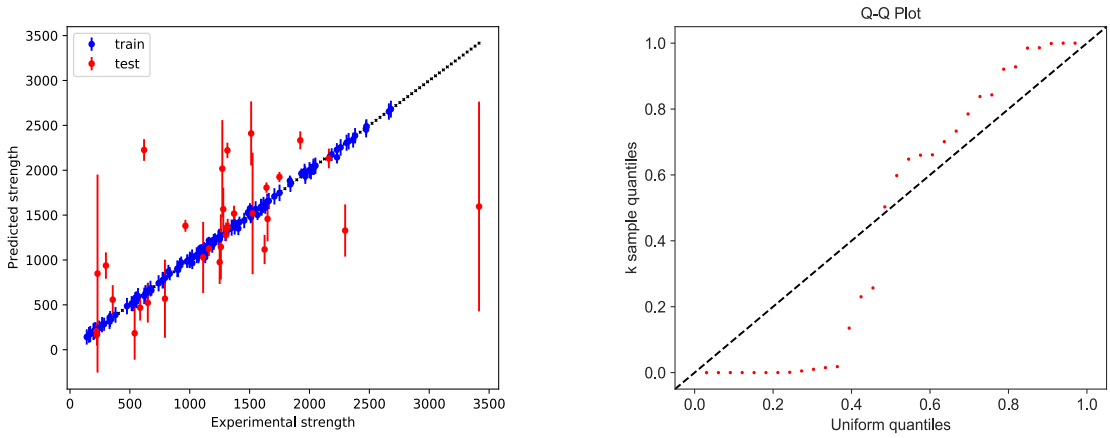
	$\text{MAE}_{\text{truth}}$	KS test statistic
SGPR	0.25	0.26
HGPR	0.12	0.11

3.2 Example 2: Yield strength prediction of HEA alloys using hardness measurements

In this example, we want to predict the expensive experimental quantity yield strength y of HEA alloy using inexpensive experimental Vickers hardness z. For doing this, we collected literature data [45]. Of the collected data, 383 HEA alloys have hardness information and 158 HEA alloys have both hardness and strength information. Using this information, the training datasets $\mathcal{D}_g = (\mathbf{X}_g, \mathbf{z}_g)$ with $N_g = 351$ hardness points and $\mathcal{D}_f = ((\mathbf{X}_f, \mathbf{z}_f), \mathbf{y}_f)$ with $N_f = 126$ hardness and strength points are constructed by keeping aside $N_v = 32$ hardness and strength points for validation. Note here, x corresponds to 25 noise-free physical descriptors (see Appendix A.1). For the sake of comparison, we built a strength model following the SGPR method utilizing the dataset $\mathcal{D}_f$. Recall that this approach assumes experimental hardness data $\mathbf{z}_f$ to be noiseless. Now following our HGPR approach, we first built a hardness model g using the data $\mathcal{D}_g$ and then from this model we estimate the de-noised hardness response $g(\mathbf{X}_f)$. Using this de-noised hardness response in the place of $\mathbf{z}_f$ in $\mathcal{D}_f$, we learned the strength response f following the procedure outlined in Sec. (2.3).

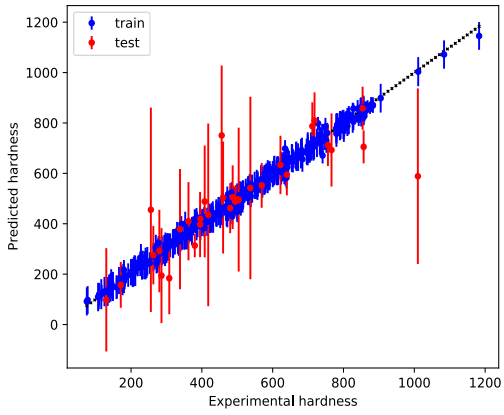
The strength predictions and q-q plots of $k_{1:N_v}$ from SGPR and HGPR methods are shown in Fig. 6 and Fig. 7, respectively. HGPR has lower MAE than SGPR. However, it is worth noting that, since the validation dataset is noisy, MAE is an unreliable measure of validation. HGPR also has better q-q plot than SGPR and the test KS test statistic is improved.

From Fig. 7a, we see that the hardness predictions at test points has room for improvement. Therefore, we decided to increase the hardness training dataset to the maximum possible size and test if this would lead to improved strength predictions. The results are shown in Fig. 8. We see a further improvement in MAE, Q-Q plots and KS test statistic. Tab. 2 shows the comparison of metrics from both the methods.

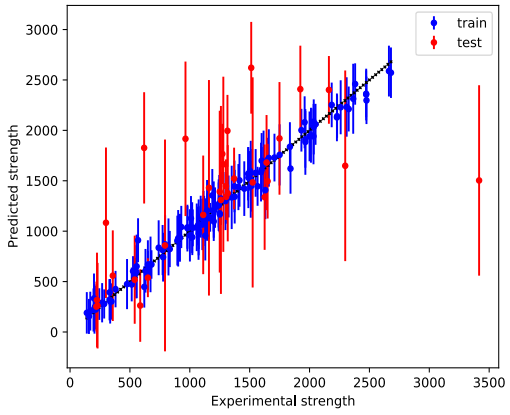


(a) (b)

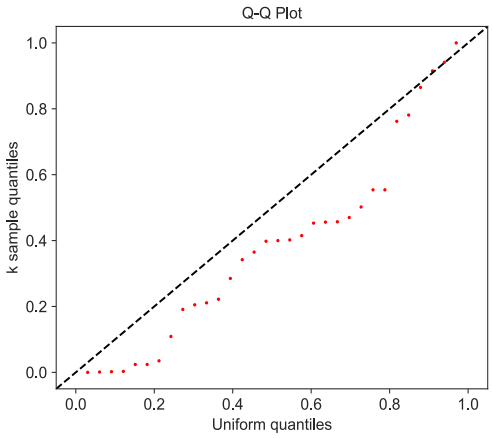
Figure 6: Example 2 - Left and right plots correspond to the predicted response and the Q-Q plot for $k_{1:N_v}$ of strength using the SGPR method, respectively. On the right plot error bars corresponds to 1.96 standard deviation band of f (MAE = 383.02, KS test statistic = 0.36).



(a)

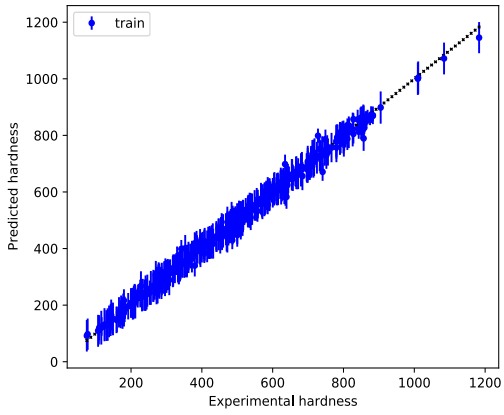


(b)

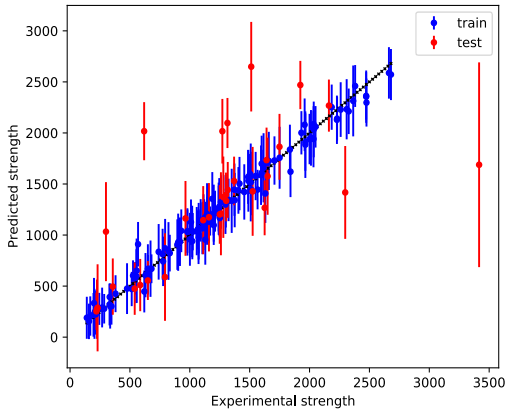


(c)

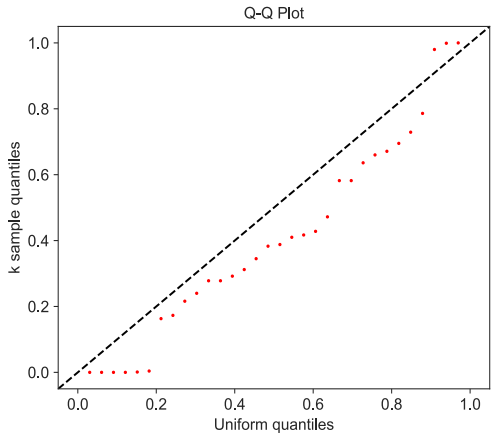
Figure 7: Example 2 - Fig. 7a and Fig. 7b correspond to the predicted response of hardness and strength from map g and f in the HGPR method, here the error bars corresponds to 1.96 standard deviation band. Fig. 7c is the the Q-Q plot for $k_{1:N_v}$ of strength (MAE = 355.48, KS test statistic = 0.26).



(a)



(b)



(c)

Figure 8: Example 2 - Fig. 8a correspond to the predicted hardness response from map g after further updating g with validation hardness data. Fig. 8b is the predicted strength from map f. Here the error bars corresponds to 1.96 standard deviation band. Fig. 8c is the Q-Q plot for $k_{1:N_v}$ of strength (MAE = 323.65, KS test statistic = 0.20).

Table 2: Example 2 - Comparison of the validation metrics from the two methods.

	MAE	KS test statistic
SGPR	383.02	0.36
HGPR	355.48	0.26
HGPR (updated with validation hardness data)	323.65	0.20

4 Conclusion

We presented a regression method denoted HGPR, that is capable of dealing with noisy inputs when one wants to correlate an inexpensive experimental measurement to an expensive one. To deal with noisy inputs, our method employs a nested model with two Gaussian processes, one going from the noiseless

physical descriptors to the inexpensive experimental measurement and other going from the noiseless physical descriptors and the inexpensive experimental measurement to the expensive experimental measurement. Towards this end, as this nested model is analytically intractable we proposed semi-analytical approximations to both the marginal likelihood and the posterior predictive distribution. We compared our method against SGPR method on a pedagogical example that demonstrates the issues of noisy inputs. Then, we applied our method to a material science application where we predict the yield strength of HEA alloys from hardness measurements. In all the cases, our HGPR method showed consistently superior performance than the conventional SGPR method. In particular, our method results in predictive distributions that better match the statistics of the data, a feature of particular importance in active learning applications.

Data availability

Data and the code to reproduce these findings will be made available upon publication of the article.

Acknowledgements

This work has been made possible by the financial support provided by the National Science Foundation (NSF) through Grant 1922316.

References

[1] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, S.-Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Advanced engineering materials 6 (5) (2004) 299–303.

[2] B. Cantor, I. Chang, P. Knight, A. Vincent, Microstructural development in equiatomic multicomponent alloys, Materials Science and Engineering: A 375 (2004) 213–218.

[3] B. Gludovatz, A. Hohenwarter, D. Catoor, E. H. Chang, E. P. George, R. O. Ritchie, A fracture-resistant high-entropy alloy for cryogenic applications, Science 345 (6201) (2014) 1153–1158.

[4] U. Bhandari, C. Zhang, C. Zeng, S. Guo, S. Yang, Computational and experimental investigation of refractory high entropy alloy mo15nb20re15ta30w20, Journal of Materials Research and Technology 9 (4) (2020) 8929–8936.

[5] Z. Li, D. Raabe, Strong and ductile non-equiatomic high-entropy alloys: design, processing, microstructure, and mechanical properties, Jom 69 (11) (2017) 2099–2106.

[6] V. K. Soni, S. Sanyal, S. K. Sinha, Phase evolution and mechanical properties of novel feconicumox high entropy alloys, Vacuum 174 (2020) 109173.

[7] Z. Li, S. Zhao, R. O. Ritchie, M. A. Meyers, Mechanical properties of high-entropy alloys with emphasis on face-centered cubic alloys, Progress in Materials Science 102 (2019) 296–345.

[8] U. Bhandari, M. R. Rafi, C. Zhang, S. Yang, Yield strength prediction of high-entropy alloys using machine learning, Materials Today Communications 26 (2021) 101871.

[9] S. Choi, S. Yi, J. Kim, B. Shin, S. Hyun, High-entropy alloys properties prediction model by using artificial neural network algorithm, Metals 11 (10) (2021) 1559.

[10] U. Bhandari, C. Zhang, C. Zeng, S. Guo, A. Adhikari, S. Yang, Deep learning-based hardness prediction of novel refractory high-entropy alloys with experimental validation, Crystals 11 (1) (2021) 46.

[11] Z. D. McClure, A. Strachan, Expanding materials selection via transfer learning for high-temperature oxide selection, JOM 73 (1) (2021) 103–115.

[12] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[13] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[14] G. E. Box, G. C. Tiao, Bayesian inference in statistical analysis, Vol. 40, John Wiley & Sons, 2011.

[15] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[16] C. E. Rasmussen, C. K. I. Williams, Gaussian process for machine learning, The MIT Press, 2006.

[17] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural network, in: International Conference on Machine Learning, PMLR, 2015, pp. 1613–1622.

[18] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, M. Bennamoun, Hands-on bayesian neural networks–a tutorial for deep learning users, arXiv preprint arXiv:2007.06823.

[19] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, arXiv preprint arXiv:2107.03342.

[20] P. Goldberg, C. Williams, C. Bishop, Regression with input-dependent noise: A gaussian process treatment, Advances in neural information processing systems 10.

[21] Q. V. Le, A. J. Smola, S. Canu, Heteroscedastic gaussian process regression, in: Proceedings of the 22nd international conference on Machine learning, 2005, pp. 489–496.

[22] K. Kersting, C. Plagemann, P. Pfaff, W. Burgard, Most likely heteroscedastic gaussian process regression, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 393–400.

[23] P. I. Frazier, A tutorial on bayesian optimization, arXiv preprint arXiv:1807.02811.

[24] G. H. Golub, C. F. Van Loan, An analysis of the total least squares problem, SIAM journal on numerical analysis 17 (6) (1980) 883–893.

[25] A. Girard, R. Murray-Smith, Learning a gaussian process model with uncertain inputs, Department of Computing Science, University of Glasgow, Tech. Rep. TR-2003-144.

[26] P. Dallaire, C. Besse, B. Chaib-Draa, Learning gaussian process models from uncertain data, in: International Conference on Neural Information Processing, Springer, 2009, pp. 433–440.

[27] A. McHutchon, C. Rasmussen, Gaussian process training with input noise, Advances in Neural Information Processing Systems 24.

[28] W. Wright, Neural network regression with input uncertainty, in: Neural Networks for Signal Processing VIII. Proceedings of the 1998 IEEE Signal Processing Society Workshop (Cat. No. 98TH8378), IEEE, 1998, pp. 284–293.

[29] G. L. Jones, Q. Qin, Markov chain monte carlo in practice, Annual Review of Statistics and Its Application 9.

[30] A. S. Weigend, H. G. Zimmermann, R. Neuneier, Clearning, in: In Neural Networks in Financial Engineering, World Scientific, 1996, p. 511–522.

[31] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian data analysis chapman & hall, CRC Texts in Statistical Science.

[32] E. Schulz, M. Speekenbrink, A. Krause, A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions, Journal of Mathematical Psychology 85 (2018) 1–16.

[33] E. Momeni, M. B. Dowlatshahi, F. Omidinasab, H. Maizir, D. J. Armaghani, Gaussian process regression technique to estimate the pile bearing capacity, Arabian Journal for Science and Engineering 45 (10) (2020) 8255–8267.

[34] A. Denzel, J. Kästner, Gaussian process regression for geometry optimization, The Journal of Chemical Physics 148 (9) (2018) 094114.

[35] D. Kong, Y. Chen, N. Li, Gaussian process regression for tool wear prediction, Mechanical systems and signal processing 104 (2018) 556–574.

[36] A. Damianou, N. D. Lawrence, Deep gaussian processes, in: Artificial intelligence and statistics, PMLR, 2013, pp. 207–215.

[37] J. S. Liu, J. S. Liu, Monte Carlo strategies in scientific computing, Vol. 10, Springer, 2001.

[38] G. Casella, C. P. Robert, Monte carlo statistical methods (1999).

[39] L. S. Bastos, A. O'Hagan, Diagnostics for gaussian process emulators, Technometrics 51 (4) (2009) 425–438.

[40] J. E. Angus, The probability integral transform and related results, SIAM review 36 (4) (1994) 652–654.

[41] Kolmogorov–Smirnov Test, Springer New York, New York, NY, 2008, pp. 283–287. doi:10.1007/978-0-387-32833-1˙214.

[42] Kolmogorov–smirnov test, See `https://www.stata.com/manuals15/rksmirnov.pdf` (Retrieved 18 June 2019.).

[43] Q-Q Plot (Quantile to Quantile Plot), Springer New York, New York, NY, 2008, pp. 437–439. doi:10.1007/978-0-387-32833-1˙331.

[44] Gpy: A gaussian process framework in python, See `https://github.com/SheffieldML/GPy` (2002).

[45] S. Gorsse, M. Nguyen, O. N. Senkov, D. B. Miracle, Database on the mechanical properties of high entropy alloys and complex concentrated alloys, Data in brief 21 (2018) 2664–2678.

A Appendix

A.1 Noiseless features employed in the construction of HEA predictive models:

There are 25 noise-free physical descriptors 'x' that were chosen based on the domain expertise in building the HEA predictive models in Sec. 3.2. The descriptors are shown in the Tab. 3 below, where c_i is the atomic percentage of the ith element, subscript i corresponds to ith element property in the alloy. Of these, the first set of descriptors were calculated based on the rule of mixtures. The next set of descriptors are estimated based on the difference between the maximum and minimum property values of the corresponding elements in the alloy. The last set of descriptors encodes the phase information of the alloy as one-hot encoding.

Table 3: Noiseless features used in the construction of HEA predictive models

Symbol	Formalism	Description
$\overline{\rho}$	$\Sigma_{i=1}^{N}\ c_i\rho_i$	Avg. Density
$\overline{Y}$	$\Sigma_{i=1}^{N}\ c_iY_i$	Avg. Young's Modulii
$\overline{T_m}$	$\Sigma_{i=1}^{N}\ c_iT_{m,i}$	Avg. Melting Temp.
$\overline{r_{at}}$	$\Sigma_{i=1}^{N}\ c_ir_{at,i}$	Avg. Atomic Radii
$\overline{G}$	$\Sigma_{i=1}^{N}\ c_iG_i$	Avg. Shear Modulii
$\overline{K}$	$\Sigma_{i=1}^{N}\ c_iK_i$	Avg. Bulk Modulii
$\overline{VEC}$	$\Sigma_{i=1}^{N}\ c_iVEC_i$	Avg. Valence e^- Conc
$\overline{\Delta S_{mix}}$	$\Sigma_{i=1}^{N}\ c_i\Delta S_{mix,i}$	Avg. Entropy of Mixing.
$\Delta\rho$	$\rho_{i,max} - \rho_{i,min}$	Range of Density
ΔY	$Y_{i,max} - Y_{i,min}$	Range of Young's Modulii
ΔT_m	$T_{m,i,max} - T_{m,i,min}$	Range of Melting Temp.
Δr_{at}	$r_{at,i,max} - r_{at,i,min}$	Range of Atomic Radii
ΔG	$G_{i,max} - G_{i,min}$	Range of Shear Modulii
ΔK	$K_{i,max} - K_{i,min}$	Range of Bulk Modulii
ΔVEC	$VEC_{i,max} - VEC_{i,min}$	Range of Valence e^- Conc.
δr_{at}	$\sqrt{\Sigma_{i=1}^{N} c_i(1 - \frac{r_{at,i}}{\overline{r_{at}}})^2}$	Asymmetry of Atomic Radii
V_{misfit}	$\Sigma_{i=1}^{N}\Delta V_i^2\ \|\|\ \Delta V = \overline{V} - V_i$	Atomic Volume Misfit
[1, 0, 0, 0, 0, 0, 0, 0]/[0, 1, 0, 0, 0, 0, 0, 0] etc.		Reduced Phase One-Hot-Encoding (O.H.E.)