A mixed categorical-continuous data-driven method for prediction and optimization of hybrid discontinuous composites performance

Raul Carreira Rufato*
ISAE-SUPAERO, Université de Toulouse, France, 31400

Youssef Diouane†
ISAE-SUPAERO, Université de Toulouse, France, 31400

Joël Henry‡ and Richard Ahlfeld§
Monolith AI, UK

Joseph Morlier¶
ICA, Université de Toulouse, ISAE-SUPAERO, MINES ALBI, UPS, INSA, CNRS, France

Surrogate models are an essential engineering tool and their popularity has increased recently due to the high computational cost of evaluating real-world simulations. However, most of these functions are described by mixed variables (continuous and categorical), which makes it harder to create accurate interpolation functions. This work builds a surrogate model from a given mixed data set, in order to quickly and accurately calculate the mechanical performance of hybrid discontinuous composites. Then, in order to find the optimal hybridization, three different approaches are performed: mono-objective, targeted and multi-objective. Starting from a virtual database provided by the industrial partner, the mixed categorical-continuous optimization process is performed by coupling a multi-armed bandit strategy and a continuous Bayesian optimization algorithm. The efficiency of our proposed approach is tested and two main results are achieved. Firstly, the obtained surrogate models are shown to be sufficiently accurate, having an R^2 score grater than 90% in average. Secondly, our proposed optimization process is able to identify correctly optimal fibres with respect to desirable targets provided by the industrial partner.

I. Introduction

TODAY, the majority of real-world simulations depends on specific categories and continuous design variables [1]. For example, a possible application is to predict the mechanical performance of hybrid discontinuous composite materials (i.e., materials manufactured by combining two or more different types of fibres). Most of these real-world black box systems can be extremely expensive to evaluate. For that reason, surrogate models are often used to accelerate the design process. Therefore, it is necessary, for engineering tasks, to build an accurate surrogate model which deals with both categorical and continuous variables and facilitates the decision-making in engineering design process.

It is commonly known to use categorical variables as outputs of machine learning models classification-based (i.e., classes). However, the work presented in this paper deals with the categorical variables as inputs to the prediction model, which often requires different methodologies. Various reviews of existing mixed continuous and categorical surrogate modeling techniques exists [2–4]. One approach [2] consists in converting the categorical variables into continuous (e.g., dummy coding) and including them in standard continuous surrogate models. Another well-known approach is based on the construction of a concept of distance between mixed variables and then, on the construction of a surrogate model through a Gaussian process [3]. A recent work [1] describes and compares the adaptation of surrogate modeling

*raulc.rufato@hotmail.com
†youssef.diouane@isae-supaero.fr
‡joel@monolithai.com
§richard@monolithai.com
¶joseph.morlier@isae-supaero.fr

techniques based on Gaussian processes for mixed continuous-categorical variables. In particular, the authors in [4] investigate the well-known kernels group, and study the generalized compound symmetry covariance matrices, that are parsimonious parameterizations of the covariance matrix, written in block form with a fixed covariance between pairs of blocks and within blocks.

The use of data-driven approach in surrogate models is a technique used when it is not possible (or expensive) to have access to the simulation code. Recently, the authors in [5] used this concept of a data-driven intelligent routine to find the best aligned discontinuous composites. The obtained results showed the potential of the proposed approach in a single and multi-objective optimization contexts. The incorporation of categorical inputs, in addition of the continuous ones, can be done using a random forest methodology [6]. However, such approach has some limitations in performing extrapolation to unknown zones of the design space. Recently, the authors in [7] proposed a method that uses a multi-armed bandit strategy to handle categorical inputs. In this proposed approach, each category corresponds to an arm with its reward distribution centered around the optimum of an objective function with continuous inputs. The obtained results [7] were very promising and show that the proposed method is able to explore well the categorical design space with a minimal computational cost. For that reason, this research chose to work with this method to perform the optimization process.

The use of machine learning (ML) algorithms in mechanical engineering has been studied a lot recently, Geetha et al. [8] presents an overview of some applications. Additionally, it has been also used to help in material design decision-making process [9]. Interestingly, the authors in [10] build a new data-driven computational framework to help design new materials and structures, building their own training set to represent the macroscopic response. Using supervised learning technics, Nyshadham et al. [11] demonstrates the utility of surrogate models to interpolate the characteristics of the atomic structure for different alloys, showing that the surrogate models can be used to select materials and their most desirable properties in a wide range of possibilities. Surrogate models are also used in [12] to identify the material parameters of ductile fracture criterion. Recent works, [13, 14] present how ML algorithms can facilitate the prediction of important material properties.

In the context of composite materials, ML approaches have been widely used [10, 11, 15–17]. Chen et al. [18] discuss and review recent progresses in ML applied to composite materials design, specifically the application of those models in the prediction of the materials' mechanical properties. Additionally, the authors in [15] train the gradient-boosted tree regression model to predict the macroscopic stiffness and yield strength of unidirectional composites. Their method is interesting as it analyses the material's micro-structure image without performing physically-based simulations. These works were able to demonstrate the excellent results in the use of machine learning models in predicting the characteristics of composite materials in order to help in engineering design. However, limitations still lie in the fact that some characteristics of hybrid composite materials can be described by categorical variables, which are usually handled by some coding method. These methods can significantly increase computational time as well as limit correlation and neighborhood characteristic analyzes between categorical and continuous variables. Our work comes precisely to contribute in this aspect. In this paper, using a mixed categorical-continuous database, we construct mixed surrogate models and then use optimization approaches to make final recommendations to help the complex engineering design process. To the best of our knowledge, this paper presents the first mixed categorical-continuous data-driven approach to predict and estimate optimal mechanical performance for hybrid discontinuous composites.

Our proposed method is used and validated in the determination of the characteristics of hybrid discontinuous composites. In fact, due to the wide variety of possible combinations of different types of fibres, it is becoming increasingly not viable and very expensive to carry out stress and deformation tests for this type of material. Thus, the proposed method for optimization and construction of accurate surrogate models is of high interest. This work, in particular, helps to optimize the scalar characteristics of the stress-strain curves for hybrid discontinuous composite materials (combination of two different fibre types). The main characteristics of this type of material can be seen in Fig. 1. The applications of composite materials are increasing in several engineering areas. Composite structures are usually chosen according to specific design goals, due to the individual constituent fibers and their dimensions and directions. Therefore, to obtain the best results, optimization models are used to find the best solutions that satisfy a given set of design constraints. [19] presents a review on the optimum design of composite structures and the relevant optimization techniques. Various optimization methods are classified according to the optimization process, but the most common approaches are conjugate directions and conjugate gradient methods. Moreover, genetic algorithms, which are commonly used methods, are part of a group of methods that do not need information about the function's gradient to find the global optimum and are generally used in optimization of composites structures [20]. For [19], genetic algorithms has been the most efficient method for obtaining the global optimum design of composites. In this paper we used another stochastic method for the optimization that was implemented by [7].

For the purpose of this work, we will first build a surrogate model (on the mixed categorical design space) and then optimize the output of the obtained surrogate. The main goal is to find the best fibre combinations according to the desired scalar parameters.

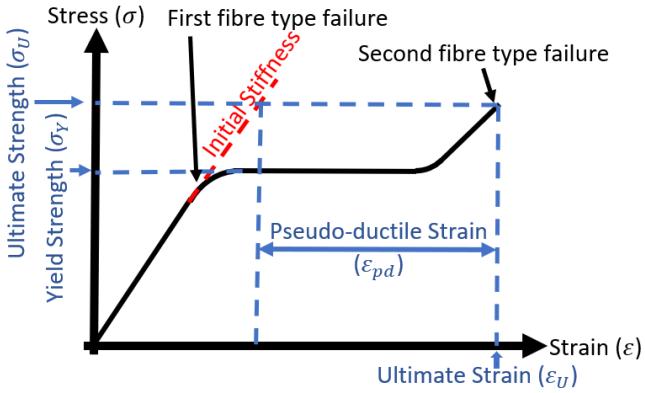


Fig. 1 A typical stress-strain curve for a hybrid discontinuous composite material.

This paper is organized as follows. Section II presents the theoretical background required by our methodology to solve the regarded problem. The application problem, with its formulation and the construction of the database, is detailed in Section III. In Section IV, we discuss the obtained results. Final conclusions are drawn in Section V.

II. Theory and methodology elements

Much of today's physical phenomena are given as expensive-to-evaluate black-box simulations. In this context, building explicit inexpensive surrogate models can be very useful to accelerate the physical modeling process. Surrogate models ought to build approximate (often polynomial) functions that interpolate the black-box simulation function and extrapolate the behaviour to unknown zones of the design space. Gaussian process are an efficient tool to build cheap and accurate surrogate models.

A. Gaussian Process

1. Gaussian Process with real inputs

Gaussian Process (GP) [21–24], also known as Kriging models, return a meta-model function f_{cont} of the form $f_{cont}(x) \sim \mathcal{N}(\mu(x), \sigma(x)^2)$ for any input $x \in \mathbb{R}^d$, where μ represents the mean (deterministic) and σ is the standard deviation that depicts the GP uncertainty of each sample on the entire domain.

A detailed description of GPs with real inputs can be as follows. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a scalar function for which a GP is built using a DoE of n points $\mathcal{D} = \{x^k, f^k\}_{k=1,\dots,n}$ where $x^k \in \mathbb{R}^d$ and $f^k = f(x^k) \in \mathbb{R}$. The GP model, related to f using n sample points, is a family of functions defined by a mean function μ and a standard deviation σ. Namely, at each point x, the GP of f is defined with a multivariate Gaussian distribution $\mathcal{N}\left(\mu(x), \sigma(x)^2\right)$. In the context of this paper, the mean μ and the standard deviation σ are given by Equation 1

$$\begin{aligned} \mu(x) &= q(x)^\top \hat{\beta} + k(x)^\top K^{-1} \left(F - Q\hat{\beta}\right) \\ \sigma(x) &= \hat{\sigma}\sqrt{1 - k(x)^\top K^{-1} k(x)}, \end{aligned} \tag{1}$$

where $q : \mathbb{R}^d \mapsto \mathbb{R}^l$ is a vector of l polynomial basis functions, $Q = [q(x^1), \dots, q(x^n)]^\top \in \mathbb{R}^{n \times l}$ is the matrix of predictors at the sampling points in $\mathcal{D}$ and $F = [f^1, \dots, f^n]^\top \in \mathbb{R}^n$ is a vector of outputs of f. For a given kernel function $k : \mathbb{R}^{d \times 2} \mapsto \mathbb{R}$, we set $k(x) = [k(x^1, x), \dots, k(x^n, x)]^\top \in \mathbb{R}^n$ as the correlation vector between the point x and the sample points of $\mathcal{D}$. K is the correlation matrix $K = [k(x^i, x^j)]_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$ computed over all the sample points of $\mathcal{D}$. Finally, the regression coefficients are defined by Equation 2

$$\hat{\beta} = \left(Q^\top K^{-1} Q\right)^{-1} Q^\top K^{-1} F \in \mathbb{R}^n, \tag{2}$$

and the the standard deviation factor $\hat{\sigma}$ of the GP is given by Equation 3

$$\hat{\sigma} = \frac{1}{n}\left(F - Q\hat{\beta}\right)^{\top} K^{-1}\left(F - Q\hat{\beta}\right). \tag{3}$$

We note that the means μ and σ are computed thanks to a correlation function chosen by the user (see the next subsection). In fact, the definition of the kernel function k may depend on a set of hyper-parameters that are in general estimated by maximizing a likelihood function.

2. Gaussian Process for mixed inputs

An efficient extension of GP to handle mixed (categorical and continuous) input variables was proposed in [1, 4]. A detailed description of the proposed methodology is as follows. Consider $f : \mathbb{R}^d \times \mathbb{F}^c \mapsto \mathbb{R}$ be a scalar function for which a GP is built using a DoE of n points of the form $\mathcal{D}_{mix} = \{x^k, z^k, f^k\}_{k=1,\dots,n}$ where $x^k \in \mathbb{R}^d$, $z^k \in \mathbb{F}^c$ and $f^k = f(x^k, z^k) \in \mathbb{R}$. The space $\mathbb{F}^c$ is a set of categorical variables $z^1, ..., z^c$ with $L_1, ..., L_c$ levels, respectively. The number of levels results in a total number of categorical combinations of $\prod_{m=1}^{c} L_m$.

The kernels on $\mathcal{D}_{mix}$ can be obtained by combining kernels for continuous and categorical variables. Let k_{cont} be the correlation kernel of continuous variable inputs and k_{cat} be the one for the categorical inputs. Different choices for the global correlation matrix k_{mix}, with respect to mixed inputs, are possible. For example, k_{mix} can be set to $k_{cont}.k_{cat}$, $k_{cont} + k_{cat}$ or to $(1 + k_{cont})(1 + k_{cat})$. In this paper, following the suggestions in [4], we set, for all $i, j = 1 \dots, n$,

$$k_{mix}(w^i, w^j) = k_{cont}(x^i, x^j)k_{cat}(z^i, z^j), \tag{4}$$

where $w^i = (x^i, z^i)$ for all $i = 1, \dots, n$. The correlation matrix for the mixed inputs is then set to be

$$K_{mix} = [k_{mix}(w^i, w^j)]_{i,j=1,\dots,n}. \tag{5}$$

For the continuous inputs, we will consider exponential kernel adapted to Gower distance correlation kernel, i.e.,for all $i, j = 1 \dots, n$,

$$k_{cont}(x^i, x^j) = \sigma^2_{cont} \exp\left(-\sum_{k=1}^{d} \theta_k \left(\frac{|x^i_k - x^j_k|}{\Delta_k}\right)^{p_k}\right), \tag{6}$$

where, for all $k = 1, \dots, d$, $\theta_k > 0$ and $1 \leq p_k \leq 2$, are hyper-parameters, σ^2_{cont} is the variance associated with the continuous kernel. For all $k = 1, \dots, d$, Δ_k is an estimation of the x_k range, i.e., $\Delta_k = (d + c)\left(\max_{i=1,\dots d} x^i_k - \min_{j=1,\dots d} x^j_k\right)$.

For the categorical inputs, the distance is measured based on a similarity score combined with the p-exponential kernel [3, 25], i.e., for all $i, j = 1 \dots, n$,

$$k_{cat}(z^i, z^j) = \sigma^2_{cat} \exp\left(-\sum_{k=1}^{c} \hat{\theta}_k \left(\frac{s(z^i_k, z^j_k)}{d + c}\right)^{\hat{p}_k}\right), \tag{7}$$

where, for all $k = 1, \dots, c$, $\hat{\theta}_k > 0$ and $1 \leq \hat{p}_k \leq 2$, are hyper-parameters, σ^2_{cat} is the variance associated to the categorical kernel. s is a score function given by

$$s(z^i_k, z^j_k) = \begin{cases} 0, & \text{if } z^i_k = z^j_k \\ 1, & \text{if } z^i_k \neq z^j_k \end{cases}. \tag{8}$$

Similarly to the continuous case, the correlation vector between a given point $w = (x, z)$ and the sample points of $\mathcal{D}_{mix} = \{w^k, f^k\}_{k=1,\dots,n}$ is given by

$$k_{mix}(w) = [k_{mix}(w^1, w), \dots, k_{mix}(w^n, w)]^{\top}. \tag{9}$$

For the sake of simplicity, in this paper, we consider to use a constant trend (zero order polynomial) for the polynomial basis functions is $q_{mix}(w) = 1 \in \mathbb{R}$. Hence, $Q_{mix} = 1_n \in \mathbb{R}^n$ a vector of size n where all the components are set to 1.

For a given $w \in \mathbb{R}^d \times \mathbb{F}^c$, the GP meta-model function $f_{mix}(w) \sim \mathcal{N}(\mu_{mix}(w), \sigma_{mix}(w)^2)$ where μ_{mix} and σ_{mix} are computed using the same formulas used for continuous inputs (given in Section II.A.1) where K, $k(x)$, $q(x)$ and

Q are replaced with K_{mix}, $k_{mix}(w)$, $q_{mix}(w)$ and Q_{mix}, respectively. We note that, as in the continuous case, the GP models for mixed inputs will require the estimation of $2(d + c) + 2$ hyper-parameters (i.e., σ_{cont}, σ_{cat}, θ_k, $\hat{\theta}_k$, p_k and $\hat{p}_k$) which is done by maximizing a likelihood function. In this paper, we considered that $p_k = \hat{p}_k = 1$ and $\sigma_{cat} = \sigma_{cont} = \sigma_0$, so only the hyper-parameters $(\sigma_0, \theta_k, \hat{\theta}_k)$ are estimated.

Now that our surrogate model has been built using a mixed training set $\mathcal{D}_{mix}$. In the next section, we will detail our proposed optimization approach to perform predictions of unknown combinations of continuous-categorical inputs, possibly not present in our initial training set $\mathcal{D}_{mix}$.

B. Bayesian optimization

Bayesian optimization [26] (BO) is an extremely useful tool to handle black-box and expensive-to-evaluate optimization problems. BO targets to find the global optimum w^* (a maximizer or a minimizer) of an unknown objective function f. Namely,

$$w^* = \arg\max_{w \in \Omega} f(w), \tag{10}$$

where Ω is the design space of interest. We note that, for simplicity reasons, we consider only a maximization formulation. A minimization problem with the objective function $-f$ can be seen also as a maximization problem.

1. Bayesian optimization for continuous inputs

The BO method with real inputs [27] can be described as follows. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a scalar function, known as the objective function. Let $\mathcal{D}^t = \{x^k, f^k\}_{k=1,\dots,t}$ be a design of experiments (DoE) of size t such that $x^k \in \mathbb{R}^d$ and $f^k = f(x^k) \in \mathbb{R}$ for all $k = 1, \dots, t$. Let μ^t and σ^t be the mean and the standard deviation of the GP associated with the DoE $\mathcal{D}^t$ (as in Section II.A.1).

BO is then a sequential search process in which, at iteration t, one enrich the DoE using a new point $(x^{t+1}, f^{t+1} = f(x^{t+1}))$. The enrichment process is led by an acquisition function $\gamma^{(t)}$ modeling the trade-off between exploration of new areas in the design space (i.e., areas with high value of σ^t) and exploitation (i.e., minimization of μ^t). Some well-known acquisition functions can be find here [27, 28]. After $t_{\max}$ objective function evaluations, the algorithm return a final recommendation $x^{t_{\max}}$. A brief description of the BO process is given in Algorithm 1.

Algorithm 1: BO for continuous inputs.

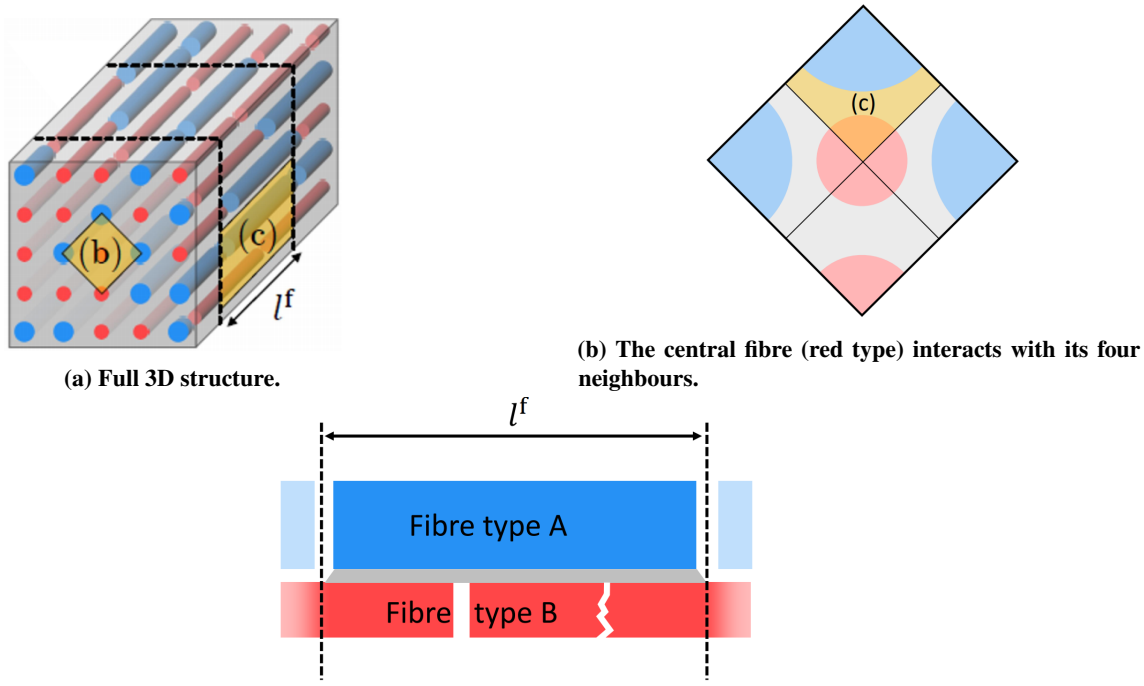
Data: : The initial DoE $\mathcal{D}^0$. A maximum number of iterations $t_{\max}$;
for $t = 0, \dots, t_{\max} - 1$ **do**
  Step 1: Build the surrogate models using GPs using the DoE $\mathcal{D}^t$;
  Step 2: Find x^{t+1} a solution of the enrichment maximization sub-problem;
  Step 3: Evaluate the objective function at $f^{t+1} = f(x^{t+1})$ and update the DoE: $\mathcal{D}^{t+1}$;
end
Return: The best point found in the final DoE $\mathcal{D}^{t_{\max}}$;

2. Bayesian optimization for mixed inputs

To handle the mixed categorical-continuous inputs, different BO based approaches can be used. One possible approach is by applying the framework given in Algorithm 1 using GP models for mixed inputs (as given in Section II.A.2). This would mean that one needs to solve the enrichment maximization problem (Step 2 of Algorithm 1) in the mixed input space $\mathbb{R}^d \times \mathbb{F}^c$. Due to the combinatorial nature of $\mathbb{F}^c$, exploring the full mixed space can be difficult, or even out of reach for large values of c and d.

In this work, we consider to use a different approach, where the enrichment maximization problem (Step 2 of Algorithm 1) will be performed only over the continuous design variables [7]. This will allow to avoid the curse of dimensionality (related to the categorical inputs) and thus the maximization problem in Step 3 will be solved with a minimal computational cost.

The regarded approach formulates the problem as a multi-armed bandit (MAB) problem [29, 30], where each category corresponds to an arm with its reward distribution centred around the optimum of the objective function over the continuous input space. Then, it identifies the best arm and maximizes the corresponding continuous function simultaneously. This is possible by considering each arm of the method as a combination of different categories. Finally, the approach uses a Thompson sampling scheme [31] to connect the MAB approach (used to explore the categorical



(a) Full 3D structure.

(b) The central fibre (red type) interacts with its four neighbours.

(c) Interaction between two types of fibres.

Fig. 2 Overview of the different scales present in the "Virtual Testing Framework" for aligned hybrid discontinuous composites [34].

space) and the Bayesian optimization approach (used to handle the continuous inputs) in a unified framework. A detailed description of the optimization is given in [7].

III. Industrial test case and overall approach

The methodology presented in Section II will be applied to a real industrial context where, using a virtual database $\mathcal{D}_{mix}$, the main goal is to build a prediction approach for hybrid discontinuous composite material. The database used in our study was generated by the "Virtual Testing Framework" [32, 33] and was created to optimize the mechanical properties of hybrid discontinuous composites [5]. The regarded framework considers a hierarchical approach where the interactions between fibres at micro scale are first considered and then used at macro level (see Fig. 2).It is out of scope of this work to explain deeply the development of this database, additional information can be obtained regarding [32, 33].

In this context, the mechanical response of interactions between fibres of different types (diameter, stiffness, strength...) was calculated using a mathematical shear lag model developed in [33] and extended to hybrid microstructures in [34, 35] (see Fig.2a). The stress-strain curves of each fibre is then computed by averaging the stress-strain curves of interactions and adding fibre breakage (Figs. 2b and 2c). In Fig. 2c, one can see that fibre breaks will occur as the stress increases. The overall stress-strain curve of the composite is then computed (Fig. 1). Finally, the final failure is calculated from a failure criterion [33] based on Dugdale's non-linear fracture mechanics criterion [36]. This Framework enables to accurately estimate the stress-strain curve of a hybrid aligned discontinuous composite of any size.

Our test database consists of approximately 10^4 points where each point of the database is described by seven input variables (2 categories and 5 continuous values) and five scalar output variables. The five continuous are described in Table 1, see also Table 2 for a sample of 5 points extracted from the tested database.

The categorical input variables are sets of different types of glass fibres and carbon fibres. We note that in this work, the words "glass" and "carbon" actually mean "low stiffness" and "high stiffness". So the same fibre can be considered as one or the other depending on what is the other fibre. The set of carbon fibres is composed of 16 different levels while the set of glass fibres is composed of 15 different ones. A detailed description of the levels is given in Table 3.

The aim of our study is twofold. First, we will build surrogate models for this hybrid discontinuous composite

Continuous inputs	
l^f	: Length of the fibres (μm)
V_c	: Proportion of carbon fibres in the composite
S	: Shear strength of the matrix (MPa)
G	: Tangent shear stiffness (MPa)
F	: Fracture toughness of the matrix (kJ/m^2)

Table 1 A detailed description of the continuous inputs.

Inputs						
Categorical z		Continuous x				
Carbon fibre	Glass fibre	l^f	V_c	S	G	F
XN-90	GF	10533.48	0.97	82.35	1056.37	0.73
XN-90	XN-05	7808.95	0.99	60.85	1741.25	0.85
XN-90	GF	9323.68	0.24	53.27	1523.66	0.67
P120J	GF	5788.76	0.45	79.63	1662.77	0.78
XN-90	XN-05	11435.55	0.84	61.73	1340.57	0.66

Outputs				
f_1: Initial Stiffness	f_2: Ultimate strain	f_3: Pseudo ductile strain	f_4: Ultimate strength	f_5: Yield strength
456482.19	0.24	0.051	864.94	864.94
466897.17	0.24	0.041	930.88	930.88
169973.97	0.18	0.033	249.34	249.34
240985.51	0.26	0.070	458.14	458.14
389829.42	0.22	0.042	693.05	693.05

Table 2 A sample of 5 points extracted from our experimental database.

materials (using a mixed database $\mathcal{D}_{mix}$), as explained in Section II.B.2. Second, we will make a final recommendation by performing 3 different mixed categorical-continuous BO optimization approaches: mono-objective, multi-objective and inverse optimization.

The overall approach proposed in this paper is outlined in the flowchart given by Fig. 3. Using a mixed database $\mathcal{D}_{mix}$, we first build a set of five surrogate models using GP. For all $i = 1, \dots, 5$, let $f_i : \mathbb{R}^d \times \mathbb{F}^c \to \mathbb{R}$ be the GP model associated with the i^{th} output. The categorical set $\mathbb{F}^c$ combines all the possible arrangements of carbon and glass fibres in our database (namely, $c = 15 \times 16 = 240$). We note that, in our database, we dispose of 5 continuous inputs, hence $d = 5$. Additionally, for practical reasons, the search space of continuous inputs $\mathbb{R}^d$ will restricted to the set $[l_b, u_b] := \{x \in \mathbb{R}^d, l_b \leq x \leq u_b\}$ where $l_b \in \mathbb{R}^d$ and $u_b \in \mathbb{R}^d$ are respectively the lower and the upper bounds. In what comes next, we will use Ω to denote the mixed targeted search space, i.e.,

$$\Omega \ := \ \left\{ w = (x, z) \in \mathbb{R}^d \times \mathbb{F}^c \ \mid \ x \in [l_b, u_b] \right\}. \tag{11}$$

Once the surrogate models are built, as outlined in Fig. 3, one can start the optimization process following 3 different approaches:

Approach I: a mono-objective approach. In this approach, we will try to maximize one desired output f_i (e.g., the initial stiffness). Namely, we select an output of interest i and perform optimization of f_i over Ω, i.e., $\max_{w \in \Omega} f(w) := f_i(w)$.

Approach II: an targeted approach. In this approach, we will try to find the set of inputs that allow to fit the best a set of targeted (or observed) output data. Namely, given a set of targeted output data $(f^*_{i_1}, \dots, f^*_{i_t}) \in \mathbb{R}^t$ where, for all

Carbon Fibres			
XN-90	P120J	T1000GB	C124
T800H	M60JB	C320	M40B
P75S	K13D	T300	XN-05
FliteStrand SZT	HTA5131	GF	C100
Glass Fibres			
GF	XN-05	FliteStrand SZT	C124
T300	T800H	C320	P75S
C100	XN-90	HTA5131	T1000GB
P120J	M40B	M60JB	

Table 3 A detailed description of the categorical sets: carbon and glass fibres.

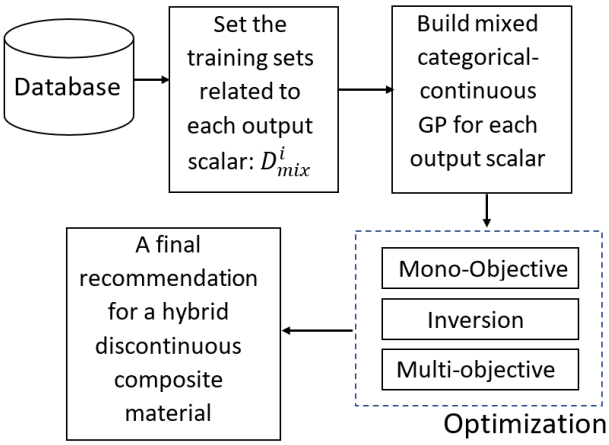


Fig. 3 An overview of the data-driven optimization process as proposed in this work.

$p \in \{1, \ldots, t\}$, $i_p \in \{1, \ldots, 5\}$ (i.e., a selected subset of the output scalars, see Table 2). Our goal is to find the input variables that minimize the misfit function. In this case, we will solve the following normalized maximization problem:

$$\max_{w \in \Omega} \; f(w) := -\sum_{p=1}^{t} \frac{\left(f_{i_p}(w) - f^*_{i_p}\right)^2}{1 + (f^*_{i_p})^2}. \tag{12}$$

Approach III: a multi-objective approach. This approach is based on a concurrent optimization process which allows experts to incorporate multiple conflicting objectives and to specify the trade-offs between them. Instead of obtaining one single optimal solution (as given by **Approach I**), a set of non-dominated solutions (known as the Pareto front) can be proposed. Using a linear scalarization approach it is possible to approximate the Pareto front. Namely, for each $\alpha := \{\alpha_i\}_{1 \le i \le t}$ such that $\{\sum_{i=1}^{t} \alpha_i = 1 \mid \alpha_i \in [0, 1]\}$, we solve the following parametrized (by α) mono-objective problem

$$\max_{w \in \Omega} f_\alpha(w) := \sum_{p=1}^{t} \alpha_p f_{i_p}(w), \tag{13}$$

where $(f_{i_1}, \ldots, f_{i_t})$ are the selected output surrogate models.

It is important to emphasize that the choice of the optimization method depends on the type of problem encountered. Therefore, the three different methods presented in Fig. 3 are not confronted, but must be chosen initially depending on the desired optimization type.

IV. Results and discussion

A. Surrogate model results

Using our database, we built mixed surrogate models for the outputs. The surrogate modeling process, as described in Section II.B.2, was implemented within the surrogate modelling toolbox (SMT) [37]. SMT is an open-source Python package consisting of libraries of surrogate modelling methods with only continuous inputs.

In order to compare the efficiency and robustness of our method, a comparison was made with 2 other existing models applied to our database. The first model used in the comparison was the one described in [5], in which a technique for encoding categorical variables called *one-hot-encoding* was used. This way of coding the categorical variables significantly increases the input dimension by one extra variable per category, because all the extra variables are set to zero except the required category which is set to one. After the codification process, this approach treats extra variables as continuous between 0 and 1. Then a Gaussian regression model of continuous variables like the one described in Section II.A.1, using the Marten 5/2 kernel, was created. The second method used for comparison, consisted of a previous codification of the categorical variables in integer values from 1 to the number of categories, thus excluding the notion of distance between categorical variables. Once the variables were coded, a *random forest* model was used to generate the surrogate model.

Models	Pearson coefficient (R^2 score)				
	f_1	f_2	f_3	f_4	f_5
Mixed GP	0.994	0.925	0.935	0.959	0.972
One Hot Marten	0.992	**0.726**	**0.899**	0.900	0.911
Random Forest	0.900	**0.831**	0.935	**0.805**	**0.860**

Table 4 A comparison between three different models: Mixed GP (our model), One Hot Marten ([5]) and a Random Forest model. In bold the scores below 90%**. (average of 10 runs)**

Due to the large size of our database, we used only 1000 points (selected randomly) to train and test the surrogate model faster. It is noteworthy that due to the increased dimension of the model when using *one-hot-encoding*, the computational training time is much higher than the other two types of model. The presented results are displayed using the average of 10 different runs. For each run, our data set was the split to two parts, 75% for training and 25% for testing the model. The obtained results for the five outputs for three different models are given in Fig. 4. For each scalar output, the Pearson coefficient R^2 is also calculated, The R^2 score measures the linear correlation between the predicted and the true values (Tab. 4 shows the R^2 score for the three different models). It is still possible to realize that using only 1000, our model was the one that got the best performance in relation to the R^2. However, one can see, from the obtained values of R^2 of our model, that the prediction accuracy (although high) was not the same for all the scalar outputs.

This can be easily explained by the stochastic nature of each output. The strain outputs (top row in Fig. 4) are much harder to predict because their variability is very high as they are a “weakest-link-based" variable: an apparently similar composite could fail at very different strains. Strength predictions (second row) are also affected by variability in the strength, although this effect is less strong. Finally, the stiffness (third row) is an "average-based" variable, and is therefore less prone to vary than failure variables. This makes it easier to predict, and explains that the accuracy of stiffness prediction is the highest. In any case, the displayed R^2 scores of our model were all above 90%, which indicates that the constructed models are sufficiently accurate.

A sensitivity analysis was performed to evaluate the impact of the inputs on the accuracy of the constructed surrogate models (f_1, f_2, ... ,f_5). The obtained R^2 scores are presented in Table 5. Our strategy here consisted of discarding a subset of inputs from the database and training the surrogate model for each situation, without such inputs. For each output, the obtained Pearson coefficient will be compared with its original value when all inputs were used.

Table 5 shows clearly that some inputs do not have a high impact on the accuracy of the prediction result. In this case, such inputs can be neglected and this, in particular, will allow later to reduce the computational cost of the optimization process (see the next subsection). In fact, by suppressing the three least influential inputs (namely, the input variables S, G and F), the Pearson coefficient for the obtained surrogate models are at worst 6.6% lower than the models built using all the inputs. As a consequence, in what comes next, we will restrict the definition of the mixed search space Ω to be of the form:

$$\Omega := \left\{ w = (x, z) \in \mathbb{R}^2 \times \mathbb{F}^{240} \mid x \in [l_b, u_b] \right\}, \tag{14}$$

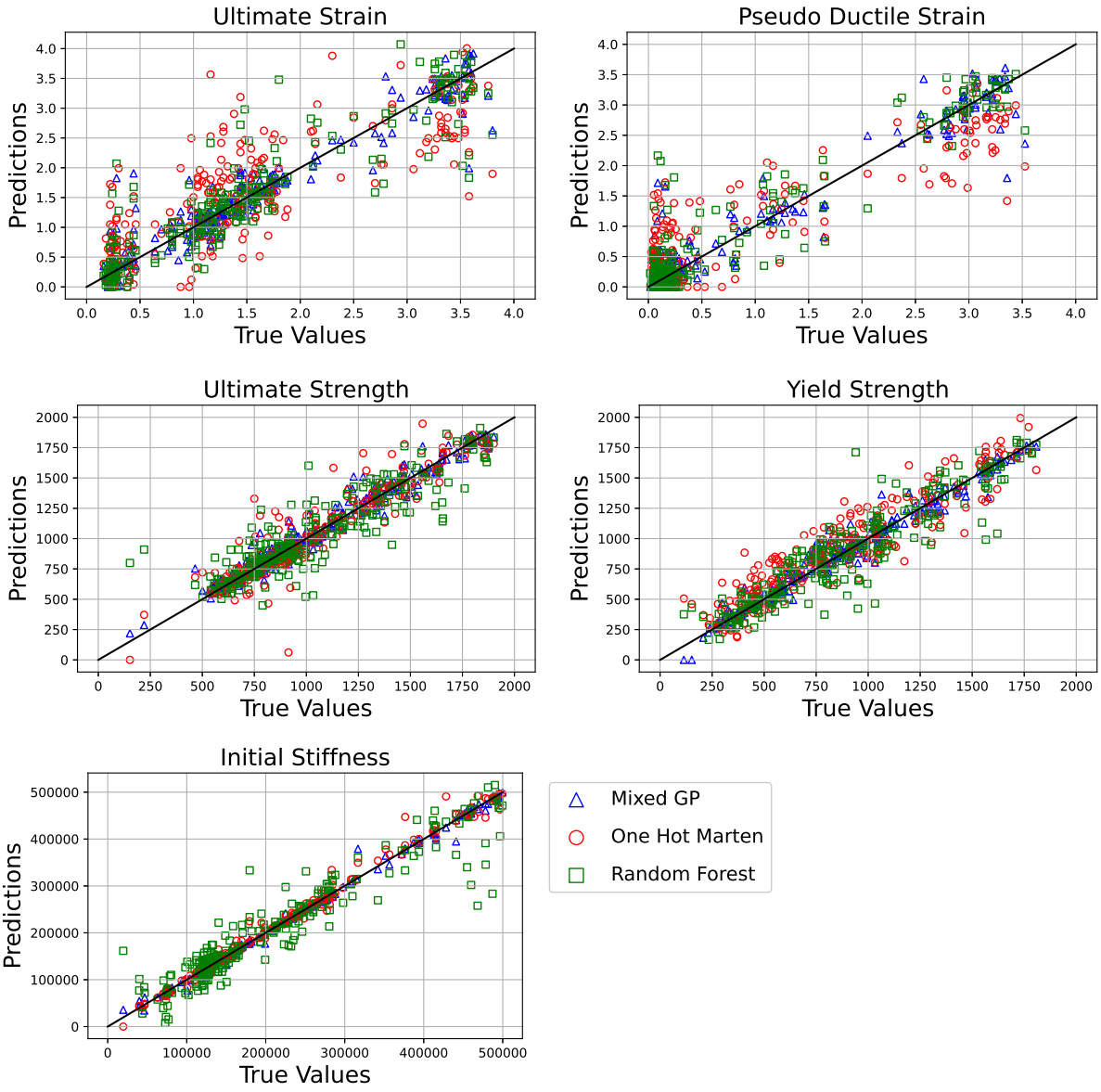


Fig. 4 Obtained results for the five scalar outputs. The black line represents the perfect scenario where the predicted values are equal to the true ones.

where $l_b \in \mathbb{R}^2$ and $u_b \in \mathbb{R}^2$.

B. Optimization results

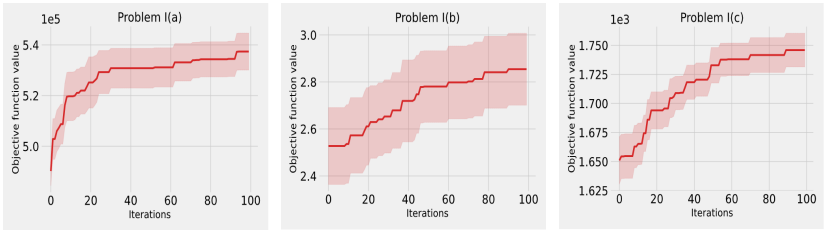
Using the trained surrogate models, we will now try to perform the optimization process following the three approaches as given in Section III. The obtained recommendations will be compared to desirable targets specified by the industrial partner "Monolith AI". Table 6 details the desirable targets and the associated optimization strategies.

For the BO solver we used the Python library Bandit-BO * which implements the approach described in Section II.B.2. All the parameters were kept unchanged, except the maximum number of iterations (set to 100), the lower bound lb (set to $[0, 515]^\top$) and the upper bound u_b (set to $[1, 12000]^\top$). Due to the stochastic natures of the Bayesian optimization solver, we performed 10 different runs and display the results in average.

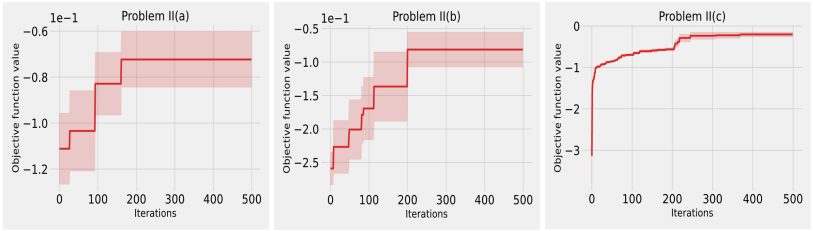
*https://github.com/nphdang/Bandit-BO

Inputs	Pearson coefficient (R^2 score)				
	f_1	f_2	f_3	f_4	f_5
All inputs	0.994	0.925	0.935	0.959	0.972
Without Carbon	**0.592**	**0.762**	**0.692**	**0.740**	**0.663**
Without Glass	**0.771**	**0.577**	**0.668**	**0.610**	**0.781**
Without l^f	0.950	**0.825**	**0.866**	**0.822**	**0.853**
Without V_c	**0.579**	**0.656**	**0.679**	**0.706**	**0.756**
Without S	0.992	0.900	0.913	0.949	0.951
Without G	0.992	0.918	0.919	0.959	0.968
Without F	0.992	0.908	0.919	0.955	0.968
Without S, G and F	0.992	**0.864**	0.901	0.911	0.902

Table 5 A sensitivity analysis using the R^2 score of the five scalar outputs with respect to the inputs. In bold the scores below 90%. (average of 10 runs)



(a) Convergence plots for problems solved using the approach I.



(b) Convergence plots for problems solved using the approach II.

Fig. 5 The obtained optimization convergence plots for problems solved using the approaches I and II. (results of 10 runs)

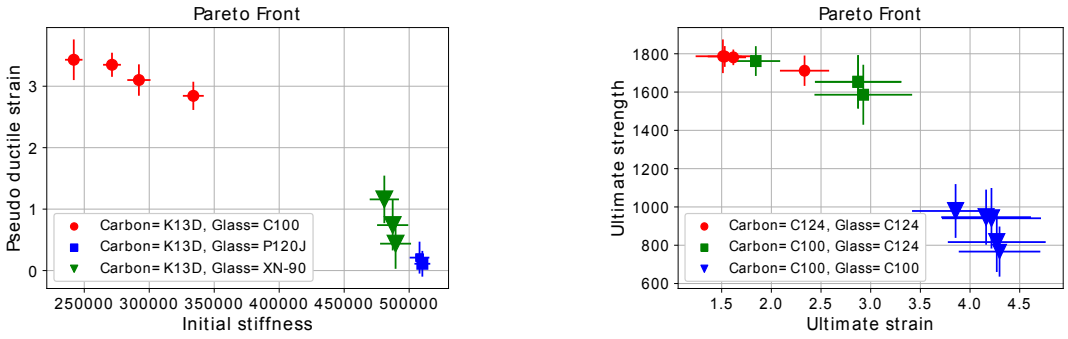
For the problems I(a), I(b) and I(c), only 100 iterations for the BO solver were used to reach convergence. For the problems II(a), II(b) and II(c), the optimization process was harder, a specific DoE was used and 500 iterations for the BO solver were used to obtain the desired convergence. The obtained convergence plots are depicted in Figure 5. Table (7) shows the obtained optimization results associated with the same optimization process. It is important to note that the Tab. (7) does not present the results for the multi-objective optimization (Problem III), since the best choice depends on the criterion adopted and the variable to which it seeks greater importance. The results for this case are presented by the Pareto front in of Fig. 6. Clearly, for all the runs, the categorical choices were almost all compatible, the expected values appeared with higher occurrence. It can be highlighted that in case II(a), K13D and XN-90 have similar behaviours and generally, when one fibre type provides good results, the other one should be good as well. Moreover, the "expected value" column was not necessarily listing all the values expected, but at least some that were expected. Then if the optimization finds other values, it should not be a problem, it just means that in that specific case, the engineer will have the choice between two different options, because they would have a similar behavior. Therefore, in case II(a) the choice could be done by the price of the fibre, choosing the cheapest one to reduce cost. The scalar values were slightly different following two scenarios. In the first one, the values were very close to the predicted range.

Problem	Strategy	Expected input values
I(a)	$\max_{\Omega} f_1$	$V_c \approx 1$ and $l^f \approx 12000$
I(b)	$\max_{\Omega} f_3$	Carbon=K13D, Glass=C100, $V_c \in [0.4, 0.6]$ and $l^f \in [3000, 5000]$
I(c)	$\max_{\Omega} f_4$	Carbon=C124, Glass=C124 and $l^f \approx 12000$
II(a)	$f_1^* = 4 \times 10^5$, $f_2^* = 1.6$, $f_3^* = 1.4$	Carbon=K13D, Glass=C124 and $l^f \approx 12000$.
II(b)	$f_1^* = 3 \times 10^5$, $f_2^* = 3.3$	Carbon=K13D or XN-90, Glass=C100 and $V_c \in [0.6, 0.7]$
II(c)	$f_1^* = 3 \times 10^5$, $f_2^* = 3$, $f_4^* = 700$	Carbon=K13D or XN-90, Glass=C100 and $V_c \in [0.6, 0.7]$
III(a)	$\max_{\Omega}(f_1, f_3)$	Carbon=K13D or XN-90, Glass=C100 and $V_c \in [0.6, 0.7]$
III(b)	$\max_{\Omega}(f_2, f_4)$	Carbon=C100,Glass=C124 and $l^f > 3000$

Table 6 The desirable targets of the industrial partner "Monolith AI" and the associated optimization strategies. The names of the problem depend on the optimization approach we consider to solve it (I, II, or III).

We believe that such values can be exactly reached using a higher number of iterations within the BO solver. The second scenario is related to outliers where possibly errors are propagated during the construction of the surrogate GP model.

In the second part of the analysis, we tried to solve problem III(a) and III(b) (see Table 6). The associate Pareto Fronts are depicted in Fig. 6. For each point of the Pareto Front we run 10 scalar optimization problems. The mean and the standard deviation are presented for each point.



(a) Problem III(a). **(b) Problem III(b).**

Fig. 6 The obtained Pareto fronts for problems solved using approach III, see Table 6. (results of 10 runs)

Due to the categorical nature of the input variables, the Pareto fronts present a clear gap. This is in particular due to the absence of intermediate mixed variables in the Pareto-optimal region. Regarding the categorical choices, clearly the two obtained Pareto front are each clustered around three (Carbon, Glass) fibres. For the different runs, the Pareto front associated with the problem III(a) (where we maximize the initial stiffness and the pseudo ductile strain, see Fig 6a) presents less uncertainties compared to the problem II(b), given in Fig 6b. Last, for both cases (Fig 6a and Fig 6b), it is interesting to note the extreme values, which prioritize more one objective than the other. In this case, the obtained results were similar to those obtained by Approach I. The analysis of the best result for the multi-objective case will depend on the criteria and needs required for the material. All the points that are on the Pareto Front are points of

Problem	Optimal Output	Optimization inputs w^*			
		Optimal V_c	Optimal l^f	Optimal Carbon	Optimal Glass
I(a)	$f_1(w^*) = 539750$	1.000	10481	K13D (10 runs)	XN-90 (5 runs) P120J (4 runs) P75S (1 runs)
I(b)	$f_3(w^*) = 2.85$	0.515	8272	K13D (4 runs) XN-90 (3 runs) C100 (3 runs)	C100 (9 runs) FliteStrand SZT (1 run)
I(c)	$f_4(w^*) = 1748$	0.147	11457	C124 (6 runs) T1000GB (3 runs) T800H (3 runs)	C124 (6 runs) T1000GB (4 runs)
II(a)	$f(w^*) = -0.072$	0.600	9736	K13D (5 runs) XN-90 (5 runs)	C124 (5 runs) GF (3 run) T300 (2 run)
II(b)	$f(w^*) = -0.081$	0.635	6462	XN-90 (6 runs) K13D (3 runs) P120J (1 run)	C100 (7 runs) T300 (2 runs) FliteStrand SZT (1 run)
II(c)	$f(w^*) = -0.203$	0.597	9575	K13D (7 runs) XN-90 (2 runs) XN-05 (1 run)	C100 (6 runs) XN-05 (3 run) P75S (1 run)

Table 7 The obtained optimization results using 10 different runs. The average of the continuous optimized values are displayed. The occurrence (out of the 10 runs) of each optimized categorical input is also specified.

optimum points, taking into consideration greater or lesser importance for the analyzed functions.

V. Conclusions

In this work, we propose a data-driven approach that deals with mixed variables types, continuous and categorical. The proposed approach was applied to an experimental mixed database related to hybrid discontinuous composite materials. Our first goal, in this work, was to build surrogate GP models. The second goal was to make a final recommendation using three different mixed categorical-continuous BO optimization approaches: mono-objective, targeted and multi-objective. The efficiency of our proposed methodology was tested. In particular, the obtained surrogate models were shown to be sufficiently accurate and our optimization process was able to identify correctly optimal fibres.

In this work, we constructed first the surrogate models and then performed the optimization of such models. It is possible to improve the optimization results by feeding directly the database to the optimization approach. This idea as well as the inclusion of extensive numerical tests will be addressed in a forthcoming work.

Acknowledgments

We would like to thanks the "Fondation ISAE-SUPAERO" that was extremly important for this publication.

References

[1] Pelamatti, J., Brevault, L., Balesdent, M., Talbi, E., and Guerin, Y., "Overview and Comparison of Gaussian Process-Based Surrogate Models for Mixed Continuous and Discrete Variables: Application on Aerospace Design Problems," *Bartz-Beielstein T., Filipič B., Korošec P., Talbi EG. (eds) High-Performance Simulation-Based Optimization. Studies in Computational Intelligence, Springer*, Vol. 833, 2019.

[2] Herrera, M., Guglielmetti, A., Xiao, M., and Coelho, R. F., "Metamodel-assisted optimization based on multiple kernel regression for mixed variables," *Struct. Multidiscip. Optimization*, Vol. 49, No. 6, 2014, p. 979–991.

[3] Halstrup, M., "Black-box optimization of mixed discrete-continuous optimization problems," Ph.D. thesis, TU Dortmund University, 2016.

[4] Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., and Wynn, H., "Group kernels for Gaussian process metamodels with categorical inputs," , 2018.

[5] Finley, J. M., Shaffer, M. S., and Pimenta, S., "Data-driven intelligent optimisation of discontinuous composites," *Composite Structures*, Vol. 243, 2020, p. 112176.

[6] Hutter, F., Hoos, H., and Leyton-Brown, K., "Sequential model-based optimization for general algorithm configuration," *International Conference on Learning and Intelligent Optimization*, 2011, p. 507–523.

[7] Nguyen, D., Gupta, S., Rana, S., Shilton, A., and Venkatesh, S., "Bayesian Optimization for Categorical and Category-Specific Continuous Inputs," *AAAI, New York, USA*, 2020, p. 507–523.

[8] Geetha, N., and Bridjesh, P., "Overview of machine learning and its adaptability in mechanical engineering," *Materials Today: Proceedings*, 2020.

[9] Huang, J. S., Liew, J. X., Ademiloye, A. S., and Liew, K. M., "Artificial Intelligence in Materials Modeling and Design," *Arch Computat Methods Eng*, 2020.

[10] Bessa, M., Bostanabad, R., Liu, Z., Hu, A., Apley, D. W., Brinson, C., Chen, W., and Liu, W. K., "A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality," *Computer Methods in Applied Mechanics and Engineering*, Vol. 320, 2017, pp. 633–667.

[11] Nyshadham, C., Rupp, M., Bekker, B., Shapeev, A. V., Mueller, T., Rosenbrock, C. W., Csányi, G., Wingate, D. W., and Hart, G. L. W., "Machine-learned multi-system surrogate models for materials prediction," *npj Computational Materials*, Vol. 5, 2019, pp. 1–51.

[12] Hou, Y., Sapanathan, T., Dumon, A., Culière, P., and Rachik, M., "A novel development of bi-level reduced surrogate model to predict ductile fracture behaviors," *Engineering Fracture Mechanics*, Vol. 188, 2018, pp. 232–249.

[13] Chang, Y.-J., Jui, C.-Y., Lee, W.-J., and Yeh, A.-C., "Prediction of the Composition and Hardness of High-Entropy Alloys by Machine Learning," *JOM*, Vol. 71, 2019, p. 3433–3442.

[14] Liu, X., Tian, S., Tao, F., Du, H., and Yu, W., *Machine learning-assisted modeling of composite materials and structures: a review*, 2021.

[15] Pathan, M. V., Ponnusami, S. A., Pathan, J., Pitisongsawat, R., Erice, B., Petrinic, N., and Tagarielli, V. L., "Predictions of the mechanical properties of unidirectional fibre composites by supervised machine learning," *npj Scientific Reports*, Vol. 9, 2019, p. 13964.

[16] Pattnaik, P., Sharma, A., Choudhary, M., Singh, V., Agarwal, P., and Kukshal, V., "Role of machine learning in the field of Fiber reinforced polymer composites: A preliminary discussion," *Materials Today: Proceedings*, 2020.

[17] Yang, C., Kim, Y., Ryu, S., and Gu, G. X., "Prediction of composite microstructure stress-strain curves using convolutional neural networks," *Materials & Design*, Vol. 189, 2020, p. 108509.

[18] Chen, C.-T., and Gu, G. X., "Machine learning for composite materials," *MRS Communications*, Vol. 9, No. 2, 2019, p. 556–566.

[19] Maalawi, K., "Introductory Chapter: An Introduction to the Optimization of Composite Structures," *Optimum Composite Structures*, Vol. 189, 2018, p. 108509. https://doi.org/10.5772/intechopen.81165.

[20] Narayana Naik, G., Gopalakrishnan, S., and Ganguli, R., "Design optimization of composites using genetic algorithms and failure mechanism based failure criterion," *Composite Structures*, Vol. 83, No. 4, 2008, pp. 354–367. https://doi.org/https://doi.org/10.1016/j.compstruct.2007.05.005, URL https://www.sciencedirect.com/science/article/pii/S0263822307001328.

[21] Abrahamsen, P., "A review of Gaussian random fields and correlation functions," *Technical report, Norwegian computing center*, 1997.

[22] Cressie, N., and Johannesson, G., "Fixed rank Kriging for very large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 70, No. 1, 2008, p. 209–226.

[23] Hensman, J., Fusi, N., , and Lawrence, N. D., "Gaussian processes for big data," *Uncertainty in Artificial Intelligence*, 2013, p. 282–290.

[24] Matheron, G., "La Théorie des Variables Régionalisées et ses Applications," *Fasicule 5 in Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris*, 1970.

[25] Gower, J. C., "Euclidean distance geometry," *The Mathematical Scientist*, Vol. 7, 1982, p. 1–14.

[26] Frazier, P. I., "A Tutorial on Bayesian Optimization," , 2018.

[27] Jones, D., Schonlau, M., and Welch, W., "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, Vol. 13, No. 4, 1998, p. 455–492.

[28] Kushner, H., "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise," *Journal of Basic Engineering*, Vol. 86(1), 1964, p. 97–106.

[29] Bubeck, S., and Cesa-Bianchi, N., "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends in Machine Learning*, Vol. 5(1), 2012, p. 1–122.

[30] Slivkins, A., "Introduction to Multi-Armed Bandits," , 2019.

[31] Thompson, W. R., "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, Vol. 25, No. 3-4, 1933, pp. 285–294.

[32] Finley, J., Henry, J., Pimenta, S., and Shaffer, M. S. P., "The influence of variability and defects on the structural performance of complex composite microstructures," *Journal of Composite Materials*, Vol. 54, 2019, pp. 565–589.

[33] Henry, J., and Pimenta, S., "Semi-analytical simulation of aligned discontinuous composites," *Composites Science and Technology*, Vol. 144, 2017, pp. 230–244.

[34] Henry, J., and Pimenta, S., "Modelling hybrid effects on the stiffness of aligned discontinuous composites with hybrid fibre-types," *Composites Science and Technology*, Vol. 152, 2017, pp. 275–289.

[35] Henry, J., and Pimenta, S., "Virtual testing framework for hybrid aligned discontinuous composites," *Composites Science and Technology*, Vol. 159, 2018, pp. 259–272.

[36] Dugdale, D., "Yielding of steel sheets containing slits," Journal of the Mechanics and Physics of Solids," *Journal of Composite Materials*, Vol. 8, No. 2, 1960, pp. 100–104.

[37] Bouhlel, M. A., Hwang, J. T., Bartoli, N., Lafage, R., Morlier, J., and Martins., J. R. R. A., "A Python surrogate modeling framework with derivatives," *Advances in Engineering Software*, 2019.