Statistical Papers (2023) 64:1251–1274
https://doi.org/10.1007/s00362-023-01437-w

REGULAR ARTICLE

Design of experiments and machine learning with application to industrial experiments

Roberto Fontana[1] · **Alberto Molena**[2] · **Luca Pegoraro**[2] · **Luigi Salmaso**[2]

Received: 29 December 2022 / Revised: 28 February 2023 / Published online: 26 March 2023

Abstract
In the context of product innovation, there is an emerging trend to use Machine Learning (ML) models with the support of Design Of Experiments (DOE). The paper aims firstly to review the most suitable designs and ML models to use jointly in an Active Learning (AL) approach; it then reviews ALPERC, a novel AL approach, and proves the validity of this method through a case study on amorphous metallic alloys, where this algorithm is used in combination with a Random Forest model.

Keywords Design of Experiments · Machine learning · Active learning · Industrial statistics

1 Introduction

In the context of product innovation, there is an emerging trend to use Machine Learning (ML) models with the support of Design Of Experiments (DOE). In this work DOE, often refers both to methods for experimental design generation and to regression models, like polynomial models. These two topics have very different backgrounds. DOE can be perceived as a classic technique, because there are industrial applications involving this topic that date back to the 1950s and earlier Bisgaard (1992), while use of ML in industry can be considered quite recent. Moreover, DOE has a precise

✉ Roberto Fontana
roberto.fontana@polito.it

Alberto Molena
alberto.molena.1@phd.unipd.it

Luca Pegoraro
luca.pegoraro.7@phd.unipd.it

Luigi Salmaso
luigi.salmaso@unipd.it

[1] Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy

[2] Department of Management and Engineering, University of Padova, Padua, Italy

and organised approach that leans on a vast and established body of literature, while ML is still mainly application-oriented. Another relevant difference between these two disciplines is the fact that while DOE tends to favour inference over predictions, allowing the experimenter to understand the existing relationships between input factors and output responses, ML models tend to behave as black boxes. Especially when the underlying phenomenon has a non-linear behaviour, the predictive performances of ML models are not always met by the traditional approaches used in DOE. What makes ML models very interesting is their ability to continuously learn and improve as more data are supplied: this characteristic matches with the principle of sequential experimentation in DOE. In ML literature, Active Learning (AL) is a kind of supervised learning technique devoted to the iterative collection of the most informative data points, with the aim of maximising information gain Olsson (2009).

From the analysis of the literature, it is possible to affirm that if we consider DOE and ML individually, their respective research areas are broad and have been intensively investigated; but things change if we consider these two topics jointly. Two different works Arboretti et al. (2022) and Freiesleben et al. (2020) state that there are few papers that consider DOE and ML jointly. Nevertheless, it is possible to identify Arboretti et al. (2022) two main currents: one concerns the utilization of ML techniques in order to analyse data that have been collected according to a DOE, and the other regards the use of DOE to optimize the training process of ML algorithms. As regards the first, in recent years the application of DOE and ML has begun to take hold, with several applications in many fields. We will delve into the analysis of this category the next section. The second current contains various contributions on the use of DOE as a method for choosing the best combination of hyperparameters for ML models: the contributions of Lujan-Moreno et al. (2018) and Staelin (2003) represent two examples of this. A systematic literature review on ML and DOE for product innovation performed by Arboretti et al. (2022c) showed that in recent years the interest on DOE+ML has grown and, in 2019 and 2020 there was a spike in the publication of papers about this topic. Moreover, Arboretti et al. (2022c) has shown that the typical application of the DOE+ML framework is non-sequential; only 8 out of the 82 analysed papers included the use of some features of the ML model to suggest the choice of the next experimental configurations.

The aim of this paper is firstly to provide a solid review of some contributions in the field of AL: to this purpose, Sect. 2 reviews a contribution about the choice of the best design and ML method for a joint application in the context of the prediction of a phenomenon of interest in physical experiments. Section 3 describes ALPERC, a recently-developed AL approach suitable for physical experiments. It is compared with other AL approaches in Sect. 4. The main novelty element of the paper is discussed in Sect. 5, that presents a real case study, in which this new AL approach is jointly used with a Random Forest (RF) model. Conclusions are in Sect. 6.

2 Experimental designs and machine learning models

In this section, the connection between experimental designs and ML models is investigated. The aim of this part is to carry out a review to understand which experimental

Table 1 Summary of the experimental designs used in the simulation study

ID	Description	# Levels	Replication
CCD	Central composite design	3	0%
BBD	Box-Behnken design	3	0%
FFD	D-optimal full factorial design	6	0%
D_opt	D-optimal design	6	0%
I_opt	I-optimal design	6	0%
LHD_rand	Random latin hypercube design	52	0%
MAXPRO	MaxPro space-filling design	52	0%
MAXPRO_dis	MaxPro discrete numeric design	6	0%
D_opt_50%repl	D-optimal design	6	50%
I_opt_50%repl	I-optimal design	6	50%
MAXPRO_dis_50%repl	MaxPro discrete numeric design	6	50%
MAXPRO_dis_25%repl	MaxPro discrete numeric design	6	25%

All designs have 6 factors and 52 runs

design is more suitable for a joint application with ML models when the global focus is on the prediction of a phenomenon of interest. In the following sections, we will analyse a study by Arboretti et al. (2022b), that considers 12 experimental designs, which will be presented in Sect. 2.1 and 7 different ML models, presented in Sect. 2.2. These will be tested considering 7 test functions (each test function is a computer simulator which models a physical phenomenon) under 8 different noise settings, including both homoschedastic and heteroschedastic noise.

2.1 Experimental designs

Table 1, contains a summary of the different DOEs settings which have been studied. It is possible to distinguish three main categories of experimental designs: "classical designs", "optimal designs" and "space-filling designs". There are six factors, with three levels each. The same number of runs (52) was allocated to each DOE, in order to provide a fair comparison between the different designs.

The data collected through the different experimental designs were used to predict the behaviour of different test functions, that are deterministic functions, mainly simulating some physical processes, described in Table 2. These test functions are detailed in the supplemental material of Arboretti et al. (2022b). The dependent variables were standardized:

$$y_n^{\text{std}} = \frac{y_n - \overline{y}}{s_y} \tag{1}$$

where y_n^{std} is the standardized value corresponding to y_n (the observed value for the n-th observation), $\overline{y}$ and s_y are the mean and standard deviation of y respectively. 100 random Latin Hypercube Designs (LHDs) with 500, 000 observations each are used for computing $\overline{y}$ and s_y.

Table 2 Summary of the test functions chosen for the simulation study

ID	Description
Borehole	Models water flow through a borehole
OTL circuit	Models an output transformerless push-pull circuit
Piston	models the circular motion of a piston within a cylinder
Piston mod	A modification of Piston, with increased non-linearity
Robot arm	Models the position of a robot arm which has 3 segments
Rosenbrock function	Is a popular test problem for optimization algorithms
Wing weight	Models a light aircraft wing

The classical designs category includes Central Composite Designs (CCDs), Box-Behnken Designs (BBDs) and Full Factorial Designs (FFDs), which are among the most used designs when it comes to data collection in ML studies. The optimal design category includes D-optimal and I-optimal designs. It is worth noting that

– both FFD and D_opt are 52-run D-optimal designs which have been generated using a 6^6 full factorial design as candidate set. The difference lies in the algorithms used for their construction;
– both D_opt and I_opt have been generated without adapting their criteria functions to take into account heteroschedasticity.

Space-filling designs, namely Random Latin Hypercube Designs (LHD_rand) and MaxPro space-filling designs (MAXPRO), are almost exclusively used in computer experiments because they have too many levels for the factors. This makes the experimentation too costly or unfeasible when it comes to physical experiments; however, these designs were included in the analysis to provide a benchmark for researchers working in computer experiments. Lastly, it is crucial to underline that in this analysis also a "hybrid" design, derived from the space-filling literature but with characteristics that enable its applications also on physical experiment, was considered. This hybrid design is the MaxPro discrete numeric design (MAXPRO_dis) Joseph et al. (2020). Also, the role of replication was investigated in the simulation study: additional D-Optimal, I-Optimal, MAXPRO_Dis designs with 50% level of replication and a MAXPRO_Dis with 25% level of replication were included. The 50% level of replication of D_opt, I_opt, and MAXPRO_dis (D_opt_50%repl, I_opt_50%repl, and MAXPRO_Dis_50%repl, respectively) have been obtained generating optimal 26-run designs and replicating them twice. This type of procedure is often performed in DOE studies concerning physical experiments. The 25% level of replication (MAXPRO_dis_25%repl) has been obtained generating a 39-run MAXPRO discrete design and randomly choosing 13 runs out of the 39 to be replicated once.

2.2 Machine learning models

As regards machine learning models, from some evidence in the literature review conducted by Arboretti et al. (2022c) it has emerged that Artificial Neural Networks

Table 3 Noise settings used in the simulation study

ID	Noise σ	Type	Description
0%	$\sigma_{hom} = 0 s_y$	–	0% noise, deterministic function
5%	$\sigma_{hom} = 0.05 s_y$	Hom	5% noise
12.5%	$\sigma_{hom} = 0.125 s_y$	Hom	12.5% noise
20%	$\sigma_{hom} = 0.2 s_y$	Hom	20% noise
50%	$\sigma_{hom} = 0.5 s_y$	Hom	50% noise
h50	$\sigma_{het,min} = 0.05 s_y$	Het.	Moderate: 5% noise at min y
	$\sigma_{het,max} = 0.5 s_y$		and 50% noise at max y
h100	$\sigma_{het,min} = 0.05 s_y$	Het.	Intermediate: 5% noise at min y
	$\sigma_{het,max} = 1 s_y$		and 100% noise at max y
h500	$\sigma_{het,min} = 0.05 s_y$	Het.	Severe: 5% noise at min y
	$\sigma_{het,max} = 5 s_y$		and 500% noise at max y

(ANNs) are the most used models for the analysis of DOE when the focus is on prediction. Two different types of ANNs were used: ANN shallow (ANN_sh), which is an ANN with one hidden layer and a number of neurons chosen in the range $[3 - 12]$ and ANN deep (ANN_dp), which is an ANN with multiple (2 to 4) hidden layers, and either 6 or 12 neurons per layer. Several other models were considered, including Support Vector Regression models (SVRs), Gaussian Processes (GPs), which are the most used when it comes to computer simulation, Linear Models (LMs) based on quadratic regression with interactions, Random Forests (RFs) and other models contained within the Automated Machine Learning (aml) platform offered by H2O LeDell and Poirier (2020). Both homoscedastic and heteroscedastic noise situations were considered (Table 3). Let s_y, min y, and max y be the standard deviation, the minimum, and the maximum of y computed using the previously mentioned 100 random LHDs, respectively. The homoschedatic case assumes noise components in the form $\epsilon \sim \mathcal{N}(0, \sigma^2_{hom})$ where $\sigma_{hom} = k s_y$ with $k = 0\%, 5\%, 12.5\%, 20\%, 50\%$; $k = 0\%$ corresponds to the deterministic case. The heteroschedatic case assumes noise components in the form $\epsilon \sim \mathcal{N}(0, \sigma^2_{het})$ where σ_{het} increases linearly with the value of $y = f(\mathbf{x})$. More specifically, at a given value $\mathbf{x}$ of the input $\sigma_{het} = 0.05 s_y + a(f(\mathbf{x}) - \min y)$ where $a = (m - 0.05) s_y / (\max y - \min y)$, $m = 50\%, 100\%, 500\%$. The minimum value of σ_{het} is $0.05 s_y$ for all cases and the maximum values of σ_{het} are $0.5 s_y$, s_y, and $5 s_y$. Each model has been implemented after a careful tuning of the hyperparameters with the objective of minimizing the Root Mean Square Error (RMSE).

2.3 Results and discussion

In the paper by Arboretti et al. (2022b) a methodology based on nonparametric permutation tests was used to evaluate the different designs and models. This approach is also described in Arboretti et al. (2014).

Table 4 Final rank of the designs for the homoscedastic noise cases

Overall ranking	Design	ID				
		0%	5%	12.5%	20%	50%
1	FFD	5	1	1	1	1
1	MAXPRO_dis	1	1	1	2	4
1	MAXPRO	2	1	1	4	1
4	I_opt	3	1	4	2	1
5	LHD_rand	3	1	6	6	8
5	BBD	6	6	4	4	4
7	D_opt	9	8	8	7	4
8	MAXPRO_dis_25%repl	7	7	7	8	8
9	MAXPRO_dis_50%repl	10	9	9	9	8
10	I_opt_50%repl	11	11	10	10	4
11	CCD	8	9	11	10	12
12	D_opt_50%repl	12	12	12	12	11

ID indicates the noise setting for each analysed scenario

Table 5 Final rank of the designs for the heteroscedastic noise cases

Overall ranking	Design	ID		
		h50	h100	h500
1	MAXPRO_dis	2	1	1
2	BBD	3	4	1
3	I_opt	3	1	6
3	MAXPRO	1	3	6
5	MAXPRO_dis_25%repl	6	4	1
6	MAXPRO_dis_50%repl	9	9	1
7	I_opt_50%repl	10	9	1
7	D_opt	8	4	8
7	FFD	5	7	8
10	LHD_rand	6	7	8
11	CCD	11	11	11
12	D_opt_50%repl	11	11	12

ID indicates the noise setting for each analysed scenario

2.3.1 Ranking of DOEs

Tables 4 and 5 report the final rankings of the experimental designs. The rankings are based on RMSE. Table 4 reports the final ranks of the designs in the homoscedastic noise settings, while Table 5 reports the final ranks of the designs in the heteroscedastic noise settings. These tables should be read column-wise because the relative ranks are computed for each noise setting. Then, by adding up all the different positions obtained by each design for all noise settings we obtained the overall ranking. For example, con-

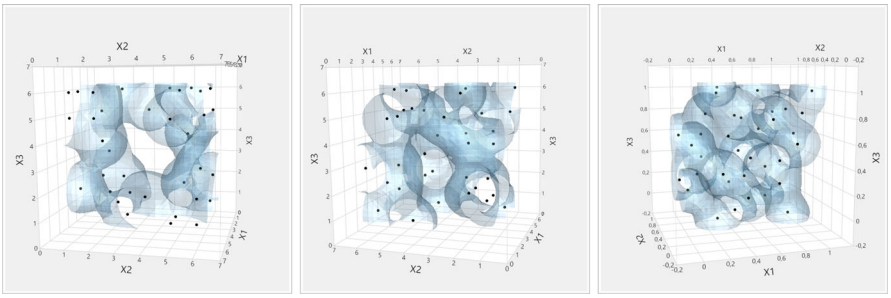

Fig. 1 Visualization of the space-filling capability of D_opt (left figure), I_opt (center figure) and MAXPRO_dis (right figure)

sidering Table 4, by adding up all the positions in the different noise settings for FFD we get 9 (it is the sum of the values in the first row); this value is the lowest obtained among all designs; that's why the position of FFD in the ranking is the first one. It is possible to observe that the choice of the experimental design has an impact on the quality of the outcome of the analysis. Focusing on the homoscedastic case, the three best ranked designs are FFD, MAXPRO_dis and MAXPRO, closely followed by the I-optimal design. The presence of replicates makes the predictions worse; D_opt_50%repl, I_opt_50%repl, MAXPRO_dis_50%repl and MAXPRO_dis_25%repl are among the worst performers. The situation is different when it comes to the heteroscedastic noise setting. The best performer is MAXPRO_dis, ranking first for the intermediate and severe noise situations, and second in the case of moderate heteroscedasticity. In this situation the two worst performers are the CCD and the D_opt_50%repl. If we jointly consider the homoscedastic and the heteroscedastic noise settings, it is possible to state that the best overall performer is the MAXPRO_dis. For this reason, we can affirm that even if a heteroscedastic noise, not expected or initially detected, appears, the best choice would be the MAXPRO_dis design, as it results among the best methods in the homoscedastic case (as it may be observed in Table 4) and the best method in the heteroscedastic case (Table 5). The results obtained by MAXPRO_dis may be justified by the fact that it uses the space-filling criterion which leads to a combination of the factor settings that maximises the ability of several different predictive models to capture the non-linearity of the underlying functions. At the same time the limited number of factors levels makes the design robust to the presence of noise and applicable for physical experiments. This design performs better than all the other designs with the same number of factor levels, FFD, D_opt and I_opt, particularly in the heteroschedastic setting. A possible explanation to this phenomenon is represented in Fig. 1, which visually compares the ability of different designs to appropriately fill the design space, while sharing the same characteristics for factor levels and runs. The space-filling criterion at the basis of the MAXPRO_dis enables a better filling of the design space favours flexible non-linear predictive models in capturing the behaviour of the underlying function across the whole experimental region.

It is worth underlining the difference, in the performances, between two classical designs: BBDs and CCDs. From this simulation study it has emerged that the BBDs performs better than the CCDs, especially in settings influenced by large noise. If we

Table 6 Final rank of the models for the homoscedastic noise cases

Overall ranking	Model	ID				
		0%	5%	12.5%	20%	50%
1	GP	1	1	1	1	1
2	LM	2	2	2	2	3
3	SVM	4	3	2	3	2
4	ANN_sh	3	3	4	4	3
5	RF	5	5	5	5	5
6	aml	6	5	5	5	6
7	ANN_dp	7	7	5	7	7

ID indicates the noise setting for each analysed scenario

Table 7 Final rank of the models for the heteroscedastic noise cases

Overall ranking	Model	ID		
		h50	h100	h500
1	SVM	1	2	1
2	GP	1	1	3
3	LM	3	4	4
4	RF	5	5	2
5	ANN_sh	4	3	6
6	aml	6	6	5
7	ANN_dp	6	7	6

ID indicates the noise setting for each analysed scenario

analyse the performances of the replicated design, it is possible to observe that these kinds of designs show some advantages only as the noise becomes larger and especially in the input dependent noise case. A possible explanation for this phenomenon is that the exploration of a smaller number of unique input configurations, in the replicated designs, weakens the ability of the predictive model to learn the behaviour of the underlying test functions; it seems that replicated design should only be preferred if the underlying phenomenon is severely affected by heteroscedasticity.

2.3.2 Predictive models

The strategy used in order to rank the different predictive models is equivalent to the one used to rank the designs, with the only difference that, in this second case, the groups are dependent since for each experimental design the same data were used in order to train the model.

From the results of the simulation, shown in Tables 6 and 7, it is evident that the choice of a specific prediction model widely impacts the results of the analysis. In the situation of homoscedastic noise, represented in Table 6, the best model is the Gaussian Process, since it ranked first in all the five cases. The performances of this model are also excellent in the situation of presence of heteroscedastic noise. This model ranks first in the low and medium noise settings and third in the high noise setting. It is also possible to state that LM is the second-best option when it comes to

situation affected by homoscedastic noise, while SVM and ANN_sh are respectively the third and the fourth options in this specific situation. RF, ANN_dp and aml behave in an unsatisfactory manner in this situation. In the case of presence of heteroscedastic noise, represented in Table 7, it is possible to observe that the SVM performs very well, indeed it is the best performer in the case of low uncertainty (together with the GP) and high uncertainty. The results obtained by LM, RF and ANN_sh can be considered as acceptable, while aml and ANN_dp perform in an unsatisfactory manner also in this situation. Lastly it is important to underline that in the simulation the focus was only on the predictive performance, and that other fundamental aspects, such as the quantification of uncertainty and the model interpretability weren't considered, even if these are two crucial factors to consider in order to obtain robust and trustworthy results that may support decision making in real industrial applications. More details about this simulation study can be found in Arboretti et al. (2022b).

3 The ALPERC method

In this section the aim is to firstly introduce some notions about AL, then to present and review the theoretical aspects of ALPERC, an iterative approach based on non parametric ranking and clustering suitable for physical experiments, recently proposed by Arboretti et al. (2022a).

3.1 Active Learning

The general framework of the AL technique requires three core ingredients: (1) an initial dataset $\Phi_0 = [\mathbf{A}_0\ \mathbf{Y}_0]$ where $\mathbf{A}_0$ is the $n_0 \times d$ matrix whose rows are the n_0 input configurations $\mathbf{x}_i = (x_{1i}, \ldots, x_{di}), i = 1, \ldots, n_0$, and $\mathbf{Y}_0$ is the $n_0 \times c$ response matrix whose rows are the vectors $\mathbf{y}_i = (y_{1i}, \ldots, y_{ci})$ of the c dependent variables corresponding to the input vector $\mathbf{x}_i, i = 1, \ldots, n_0$. (2) c predictive models, developed on the dataset Φ_0 and lastly (3) a criterion that uses some features of the model to propose which experimental configurations should be added to the dataset at the subsequent iterations. When this configuration is defined and added to the dataset, the above described steps are iterated until a stopping condition is reached.

The idea underlying this process is that by collecting data on the most informative input configurations it is possible to achieve the goal of the study more efficiently in terms of time and required resources.

Among the first algorithms proposed for the emulation of complex functions by sequential data acquisition there are the active learning MacKay (ALM) and the active learning Cohn (ALC). ALM, Yue et al. (2021), adds the input configurations which are characterised by the highest predictive uncertainty, maximising the expected information gain; two important features of this algorithm favoured its wide diffusion: it is intuitive and easy to be implemented. On the other hand, ALC, Gramacy and Lee (2009) proposes for inclusion those data points that minimize the expected integrated variance over the entirety of the input space. This results in the selection of those $\mathbf{x}'$ points that maximise the expected reduction in predictive uncertainty in the input space as in the formula:

$$\int [\sigma^2(\mathbf{x}) - \sigma^2_{\mathbf{x}'}(\mathbf{x})] d\mathbf{x}$$

where $\sigma^2(\mathbf{x})$ represents the estimated variance in $\mathbf{x}$ given the currently available observations and $\sigma^2_{\mathbf{x}'}(\mathbf{x})$ is the expected predictive variance in $\mathbf{x}$ when the configuration $\mathbf{x}'$ is included. Other approaches are proposed in the literature, like the one in Binois et al. (2019). One common characteristic between all the proposed AL criteria is that they all exploit a quantification of the predictions of the uncertainty to select subsequent experimental configurations. Another common trait between all the analysed criteria is that they deal with the analysis of computer experiments, while in this article the interest is on physical experiments.

3.2 ALPERC

Arboretti et al. (2022a) propose ALPERC, an AL approach which is based on nonparametric Ranking and Clustering and is suitable in Physical Experiments. ALPERC can be implemented for sequential data collection when three or more response variables are investigated in the same experiment in noisy settings. This algorithm is based on the combination of different building blocks:

1. an experimental design for collection of data at the first iteration and a set of candidate points from which it is possible to choose new input configurations in the subsequent iterations;
2. a predictive model developed on the available data which provides a quantification of the uncertainty of candidate points;
3. a variable importance technique;
4. a ranking procedure to obtain an inferential rank of candidate configurations concerning predictive uncertainty;
5. a clustering procedure that groups candidate configurations with respect to continguity in the design space.

The underlying idea is to propose a model-agnostic AL methodology, with the only strict condition that the predictive models must provide an appropriate quantification of uncertainty of predictions. In Arboretti et al. (2022a) the focus was on Gaussian Process models, as they are the most common in the AL literature and they can directly provide a quantification of uncertainty, but other models can also be used, as we will see in the case study. As regards the variable importance, the choice of the appropriate technique depends on the predictive model selected. In case Gaussian Process models are selected, an appropriate strategy consists of the use of the Sobol' indices. These consider independent input variables and quantify the relative importance of one input dimension as the partial variance of model output explained by this variableWei et al. (2015).

Let's consider an objective function $f(\mathbf{x})$, that is assumed to be square-integrable; Sobol's method considers the functional decomposition of $f(\mathbf{x})$:

$$f(x_1, ..., x_d) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i=1}^{d} \sum_{j>i}^{d} f_{ij}(x_i, x_j) + \cdots + f_{1,...,d}(x_1, ..., x_d) \quad (2)$$

where f_0 is a costant that represents the mean value of $f(\mathbf{x})$, $f_i(x_i)$ is the main effects of x_i, $f_{ij}(x_i, x_j)$ is the interaction effect between two different factors x_i and x_j, and $f_{i_1 \ldots i_k}(x_{i_1}, \ldots, x_{i_k})$ is the interaction effect among the factors $x_{i_1}, \ldots x_{i_k}$, $k > 2$, and $i_1 < \ldots < i_k$.

Sobol demonstrates that if the input variables are independent and $f(\cdot)$ is square integrable, from Eq.(2) the variance associated to the model response Y can be written as:

$$V(Y) = \sum_{i=1}^{d} V_i + \sum_{i=1}^{d} \sum_{j>i}^{d} V_{ij} + \cdots + V_{1,\ldots,d} \tag{3}$$

where $V_i = V(\mathbb{E}(Y|x_i))$, $V_{ij} = V(\mathbb{E}(Y|x_i, x_j)) - V_i - V_j$ and so on.

The first-order sensitivity indices S_i are expressed as:

$$S_i = \frac{V_i}{V(Y)} = \frac{V(\mathbb{E}(Y|x_i))}{V(Y)}, \ i = 1, \ldots, d \tag{4}$$

The index S_i measures the proportion of variability in the response that is attributable to the i-th input variable. Another relevant indicator is the total partial variance V_{T_i} associated to the i-th input variable, as it considers not only the main effect of the i-th input, but also its interaction effects with all the other $d-1$ input variables $\mathbf{x}_{\sim i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. From Eq.(3), $V_{T_i} = V(Y) - V(\mathbb{E}(Y|\mathbf{x}_{\sim i}))$ is obtained. The total sensitivity index S_i can be computed as:

$$S_{Ti} = \frac{V_{T_i}}{V(Y)} = \frac{V(Y) - V(\mathbb{E}(Y|\mathbf{x}_{\sim i}))}{V(Y)} \tag{5}$$

In practical applications Sobol' indices can be obtained by Monte Carlo simulations. Sobol (2001)

The problem of clustering, which is an unsupervised classification task, is related to the grouping of objects based on some measure of similarity. In Arboretti et al. (2022a) a hierarchical agglomerative clustering algorithm based on weighted Euclidean distance was proposed. The choice of this criterion was made because this algorithm results more stable than other common choices, such as k-means clustering, as it is insensitive to the initial seed selection, ensuring replicable results. A centroid-linkage was considered and in order to estimate the similarity between two vectors $\mathbf{x}$ and $\mathbf{x}'$ the weighted Euclidean distance was used:

$$\left(\sum_{i=1}^{d} w_i (x_i - x_i')^2 \right)^{1/2} \tag{6}$$

where w=(w_1,…,w_d) represent the vector of the weights assigned to the d dimensions. To identify the best number of clusters the Silhouette index was chosen, as it has been shown to be the best methods in most situations. Arbelaitz et al. (2013)

3.3 How ALPERC works

The algorithm begins by considering an initial dataset $\Phi_0 = [\mathbf{A}_0\ \mathbf{Y}_0]$, where $\mathbf{A}_0$ is often chosen in the class of Maximum Projection design with discrete numeric factors and $\mathbf{Y}_0$, is the matrix of responses, assumed to be independent. Then an additional design $\mathbf{A}_{cand}$ that includes n_{cand} candidate configurations is built. The number of n_{cand} have to guarantee an appropriate coverage of the design space, however it should be remembered that the computational effort increases with the size of $\mathbf{A}_{cand}$, thus a valid option may consist of the temporary augmentation of $\mathbf{A}_0$ at each AL iteration by the selection from $\mathbf{A}_{cand}$ of a limited number of combinations which minimize the Maximum Projection criterion, and to use this subset as a candidate set at that specific AL iteration. A commonly used value for the number of candidate points at each iteration is 100.

The next steps of the procedure are the construction of c predictive models $f_1(\cdot), \ldots, f_c(\cdot)$ that are trained on Φ_0 and the quantification of the importance of each feature $x_i,\ i = 1, \ldots, d$. As already mentioned, in Arboretti et al. (2022a), a GP model was considered and the total Sobol' indices were used in order to obtain a quantification of the uncertainty for the candidates in $\mathbf{A}_{cand}$.

Then, the ranking procedure introduced by Arboretti et al. (2014), is employed, where for ALPERC the groups G considered in the paper are the rows of $\mathbf{A}_{cand}$ and the dependent variables are the predictive uncertainty observed on the c responses. The permutation tests use the difference in means as test statistics and consider c observations in each group G. Therefore, a value $c \geq 3$ is recommended to guarantee a minimal sample size for the permutation tests. As a result, the procedure provides a synthetic rank of candidates with respect to the uncertainty associated to all the responses. In comparison to the application of ALM method, the main advantage of ALPERC is that the ranks assign different positions only to those candidates whose global predictive uncertainty is significantly different after the execution of permutation tests. It may happen that some area of the design space is characterized by the highest uncertainty, so the candidates from that region will rank high. If the predictive uncertainty is the only indicator driving AL acquisition, the proposed points would all be close together and located in that specific region of the design space, but in general, one prefers to explore several areas of the experimental domain, in order to increase the global accuracy. For this reason, a clustering method is employed to group together candidates that are close together in the design space. The weighted Euclidean distance is used as a similarity measure for the generation of clusters, where the weights are given by the rescaled relative importances of each variable. The adoption of this strategy increases the possibility to put two candidates that differ with respect to "irrelevant" dimensions in the same cluster, and in an opposite way two configurations that are spatially near but differ with respect to a few decisive factors tend to be assigned to different clusters. At this point, a batch of experimental configurations must be selected from the candidate set for inclusion in the next AL iteration. First, the size of the batch n_{add} has to be set, and this is usually application-specific. A general guideline is to set n_{cand} at least one order of magnitude larger than n_{add}, in order to provide a reasonably large set of candidates at each AL iteration. Two different rules to guide sequential candidate

selection can now be chosen, one favoring the *exploration* of the design space, and the other, more conservative one, that favors *exploitation* of the current knowledge.

Let's consider a situation in which two candidate experimental configurations both share the highest rank position and are in the same cluster. In the case of *exploration* of the design space, only one of these candidate configurations is selected (the one with highest mean uncertainty) and then the rank is descended until a new candidate is found in another cluster and/or has a different position in the rank. In the case of *exploitation*, the idea is to perform some replicates of the experimental configuration characterized by the highest mean uncertainty, while sharing the same ranking position and cluster with others. The number of required replicates should be equal to the number of candidates which share the same position in the ranking and cluster. In practice, the *exploration* strategy is preferable in most situations, but in presence of severe heteroscedasticity the predictive models greatly benefit from the execution of replicates, as a separation of noise from signal can be achieved more easily. In the end, n_{add} configurations are selected from $\mathbf{A}_{cand}$ in accordance to one of the principles already explained, and the new runs can be executed. Once the new responses are collected, the new dataset $\Phi_{add,0}$ is concatenated to Φ_0 and the procedure can be iterated n_{iter} times, i.e. until a certain accuracy threshold is reached or until the company has exhausted the resources allocated for the project.

4 Simulation study: ALPERC vs competitors

In this section the aim is to review a simulation study about different AL algorithms, including ALPERC and a non-active-learning (non-AL) approach, in order to understand the validity of the AL approaches and more specifically of ALPERC. In order to achieve this goal, we have analysed the simulation study conducted in Arboretti et al. (2022a) that compares the predictive accuracy of ALPERC against some competitors from the literature. This simulation is based on several test functions, that are those already presented in Table 2, together with multiple noise settings, considering both homoscedastic and heteroscedastic situations, and two different sparsity levels: 0% sparsity and 25% sparsity, where sparsity is the ratio of the number of inactive factors with the number of factors which are considered in the experiment. As regards ALPERC, the exploration strategy was preferred in all noise situations, except the one with the highest heteroscedasticity. The others sequential data acquisition techniques which have been analysed are:

- a variation of ALPERC (ALPERC_unw), that considered all the weights equal 1, **w=1**, so the attribution of each candidate configuration to a cluster wasn't adjusted by variable importance;
- a selection based on optimisation of the Maximum Projection criteria (MaxPro_aug);
- an augmentation based on D-optimality (D_opt);
- an iterative data acquisition based on the principle of maximum variance (ALM);
- a sequential sampling based on the expected variance reduction throughout the design space (ALC);

Table 8 Rank of the different methods in the homoscedastic noise setting

Overall ranking	Method	0% sparsity		25% sparsity	
		20%	50%	20%	50%
1	ALPERC	1	1	1	2
2	MaxPro_aug	1	1	3	5
3	ALPERC_unw	6	1	4	1
4	ALC	4	6	2	2
5	D_opt	1	5	5	5
6	ALM	5	4	6	2
7	non-AL	7	7	6	7

– a non-active-learning approach in which the models are retrained at each iteration on a new design of suitable size (non-AL). A new Maximum Projection design with a limited number of levels was built at each iteration and its size that matched the other AL counterparts. This is the reference approach that shows the performance of non-sequential methods, which are, as previously seen, the most employed in the literature on DOE+ML.

From the simulation it emerges that for the homoscedastic noise setting ALPERC performs as the best method in three out of four situations, and as the second best method in the remaining one, as it is possible to observe in Table 8. In the remaining case, the one with the highest levels of uncertainty and sparsity, the best method results the ALPERC_unw. It is also important to underline that the performances of the MaxPro_aug are equal to the ones of ALPERC, when the level of sparsity is low, while, if the level of sparsity increases the perfomances of this method decreases.

As regards the heteroscedastic noise settings ALPERC always ranks first, regardless of the sparsity and noise levels, as it is possible to observe from Table 9.The unweighted version of ALPERC always ranks second, except in the case of high sparsity level and high heteroscedastic noise, when it matches the results of ALPERC. ALPERC was the only strategy including replicates at the most severe level of heteroscedasticity, because of the *exploitation* approach. In Arboretti et al. (2022a) it is underlined that even if at each AL step, the experimental configuration selected by the closest competitors had been replicated three times, to match the level of replication of ALPERC, this strategy would perform worse than ALPERC. This demonstrates that at the highest level of uncertainty, the benefits provided by the ALPERC methodology don't exclusively depend on the presence of replicates, but are a result of the essential principles of the methodology. ALPERC is a sequential algorithm that allows to reassess the situation at each iteration, so if a heteroscedastic noise not initially expected or detected appears, the most obvious choice would be to favour the *exploitation* strategy. Lastly, it is possible to state that the results of the non-active-learning approach (non-AL) can be considered as unsatisfactory, because this methodology ranks last in all noise settings.

Table 9 Rank of the different methods in the heteroscedastic noise setting

Overall ranking	Method	0% sparsity		25% sparsity	
		h100	h500	h100	h500
1	ALPERC	1	1	1	1
2	ALPERC_unw	2	2	2	1
3	D_opt	2	2	2	4
4	MaxPro_aug	2	2	2	5
5	ALM	2	2	6	3
6	ALC	2	6	2	7
7	non-AL	7	7	7	6

5 Case study: amorphous metallic alloys

5.1 Overview

This section presents a case study about the costruction and refinement of a multi-response emulator to estimate three critical temperatures in some innovative metallic alloys. To achieve the desired goal, data from real experiments are used, along with ALPERC, which results useful in the sequential data collection for iteratively refining the predictive algorithms.

As already mentioned, the case study is about innovative metallic alloys: the amorphous metals. These materials mantain, even at solid state, the typical disordered structure of the liquid state, so they don't have a cristalline structure, and for this reason they are also known as metallic glasses. The particular structure of these materials results in some very interesting properties, like high strength and wear resistance Jafary-Zadeh et al. (2018), high hardness and elasticity Chan and Sort (2015), high magnetic permeability Khan et al. (2018) and high corrosion resistance Nair and Priyadarshini (2016). Moreover, the unique characteristics of these metallic glasses make them interesting for different applications in various industries such as sporting good, advanced aerospace applications and medical and electronic devices Chan and Sort (2015).

A limit to the pratical application of these materials is caused by the fact that it is difficult to obtain amorphous alloys with a thickness greater than 1 mm. Another limit is represented by the high number of elements that are necessary to achieve an appropriate alloy structure; this also makes the size of the combinatorial space prohibitive.

The process of solidification results critical in obtaining the desired structural features of the material and the cooling process is governed by some critical transformation temperatures (CTTs):

- the glass transition temperature T_g;
- the onset of crystalizzation temperature T_x;
- the liquidus temperature T_l.

A rapid and precise prediction of the CTTs of candidate material is required to improve the properties of amorphous metallic alloys. That's why this case study regards the

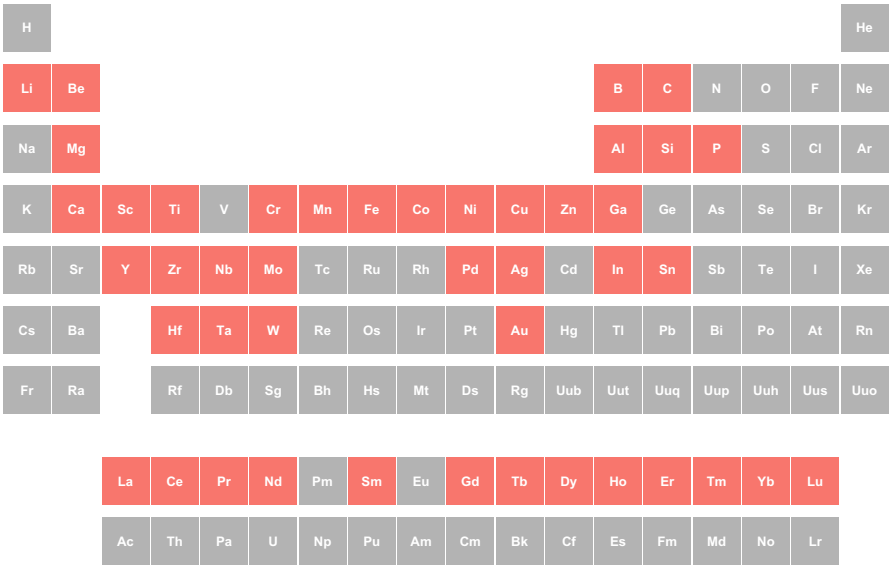

Fig. 2 The elements forming the alloys in the metallic glasses dataset

construction and the refinement of predictive models to emulate the three CTTs given both the alloy elements and the composition.

5.2 Dataset and ALPERC implementation

In the case study the data collected by Xiong et al. (2020) are employed. After the cleaning phase, the dataset consists of 555 measurements from differential thermal analysis or differential scanning calorimetry at a constant heating rate. The alloys investigated in the dataset include 44 elements, as highlighted in Fig. 2. All the observations were rescaled to $[0-1]$, using

$$z_n = \frac{y_n - y_{min}}{y_{max} - y_{min}} \tag{7}$$

where z_n is the rescaled value corresponding to y_n (the observed value for the n-th observation), y_{min} represents the minimum value in the dataset, while y_{max} represents the maximum value in the dataset.

This rescaling was performed to allow a fair comparison between variables that may have different orders of magnitude. Both the responses (T_g, T_x, and T_l) and the 44 explanatory variables have been rescaled.

It is important to underline that the data used in this case study are unstructured.

A random partitioning of the experimental data into training and test sets (80% and 20% of the data respectively) is operated, and repeated 10 times for robustness. ALPERC is applied to each initial dataset, with the following starting conditions: $n_0 = 40$, $n_{cand} = 100$, $n_{add} = 10$, $n_{iter} = 20$ and the exploration strategy. $\mathbf{A}_0$ is composed by observations randomly selected from the training data. A random design

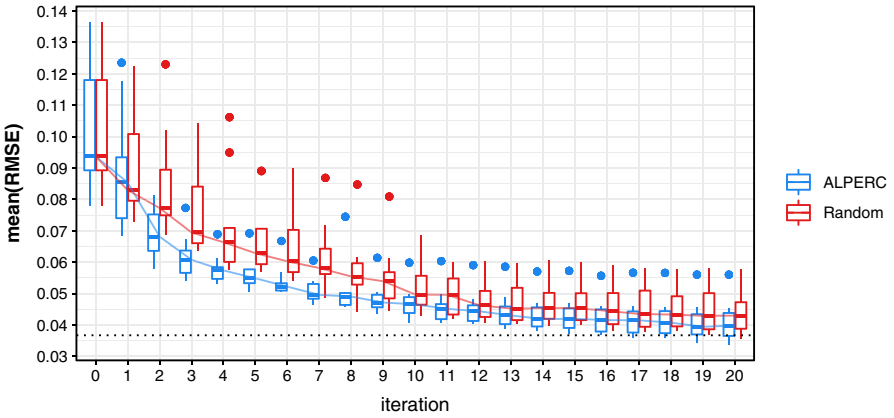

Fig. 3 Results of the average MSE at each AL iteraction. The black dotted line corresponds to the median mean(RMSE) when all training data are used for training the models

$\mathbf{A}_0$ has been chosen to focus on the behaviour of ALPERC in the active learning phase when the starting design is non-optimal, i.e. potentially worse than it could be. At each AL iteration, $\mathbf{A}_{cand}$ is constructed including the n_{cand} data configurations that optimize the Maximum Projection criterion given the experimental trials already included in the design. Considering the large number of predictors, a RF model was chosen and it has been trained using 5-fold Cross Validation Gareth et al. (2013). The uncertainty quantification associated with this ML model used in the case study follows the methodology of Wager et al. (2014). Lastly the estimation of the variable importance is performed with the permutation method Breiman (2001).

5.3 Results and discussion

The evolution of the mean test error obtained with ALPERC is represented in Fig. 3. From the comparison with the baseline approach, it appears that ALPERC is preferable to the random sampling of candidate configurations. The median test error obtained when using 85% more data than ALPERC at iteration 20 is represented in Fig. 3 by the black dotted line: this is another proof of the good predictive accuracy obtained by the models using ALPERC.

As it is possible to observe from the three scatterplots in Fig. 4, which represent the observed vs predicted values, the accuracy on the test data is very high for all the three responses.

In Fig. 5 the evolution of the variable importance of the T_g through ALPERC iterations is represented. To improve the visual impact, those variables for which the median variable importance always results smaller than 10% in each of the AL iterations and for all the CTTs are displayed in light grey. From this plot emerges that 29 out of the 44 predictors are barely important for all the responses or, equivalently, that only 15 predictors should be taken into account. This is a very useful information, because it allows to understand which elements need the most investigation.

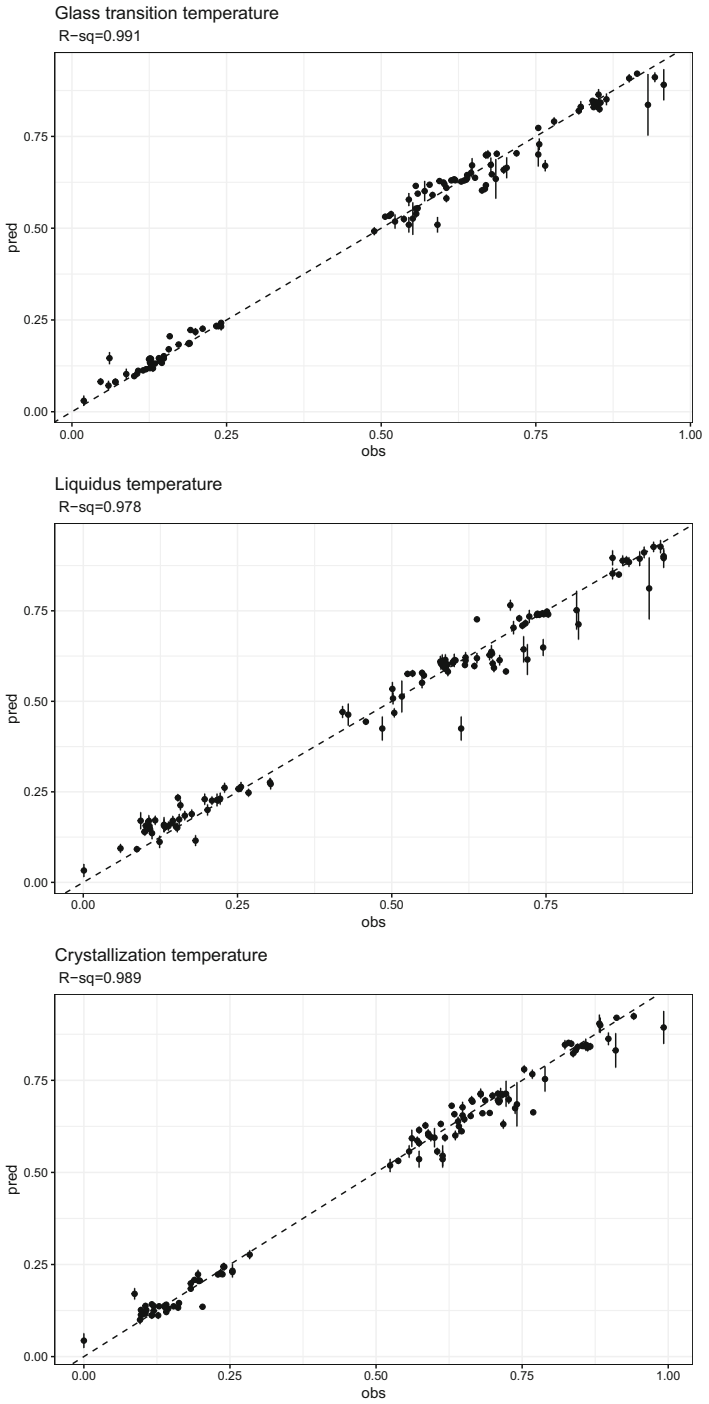

Fig. 4 Scatterplot of observed vs predicted values on the test data, considering the training-test partion that leads to the best results at iteration=20 for ALPERC

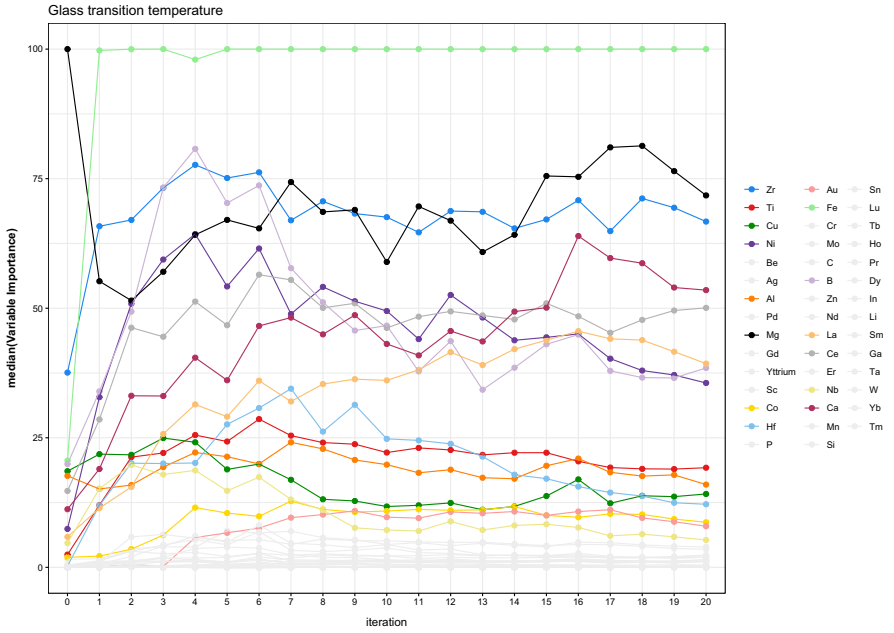

Fig. 5 Median variable importance over the ALPERC iterations for T_g

To better comprehend the RF model employed, the SHAP (SHapley Additive exPlanations) techniqueLundberg et al. (2018) can be used. The SHAP values are inspired from the Shapley indices of game theory literature Shapley (2016), and provide an explanation of individual predictions.

By calculating the contribution of each feature to the prediction, SHAP aims to explain the prediction of a single instance **x**: this can be achieved by assuming an additive model form and analysing, by using the Shapley values, how much each variable affects the prediction for the instance **x** in relation to the overall mean prediction calculated on a given dataset.

Let's consider the response T_g as an example (Fig. 6): the plot indicates that the average test data prediction is 0.496 (for each response the values have been rescaled to $[0-1]$). Then, the value of the predictor Co$=0.4$ adds 0.061 to the mean prediction, while the content of Zirconium, Zr$=0$, subtracts 0.054 to the mean prediction, and so on. Considering all the input variable values, the final prediction for the selected test instance is of 0.509, corresponding to 626.63K. To aid in visualization, just the contribution of the five most relevant predictors is shown for the given test instance, whereas the rest is collapsed in the "other variables" category. For the other responses this approach leads to a final prediction of $T_x = 682.72$K and $T_l = 996.65$K. This visualization offers a precise explanation of how each predictor contributes to the prediction of a certain data configuration, providing also insights on the rationale underlying the ML model.

By plotting the feature value associated with each test instance on the horizontal axis and the related SHAP value Molnar et al. (2020) on the vertical axis it is possible

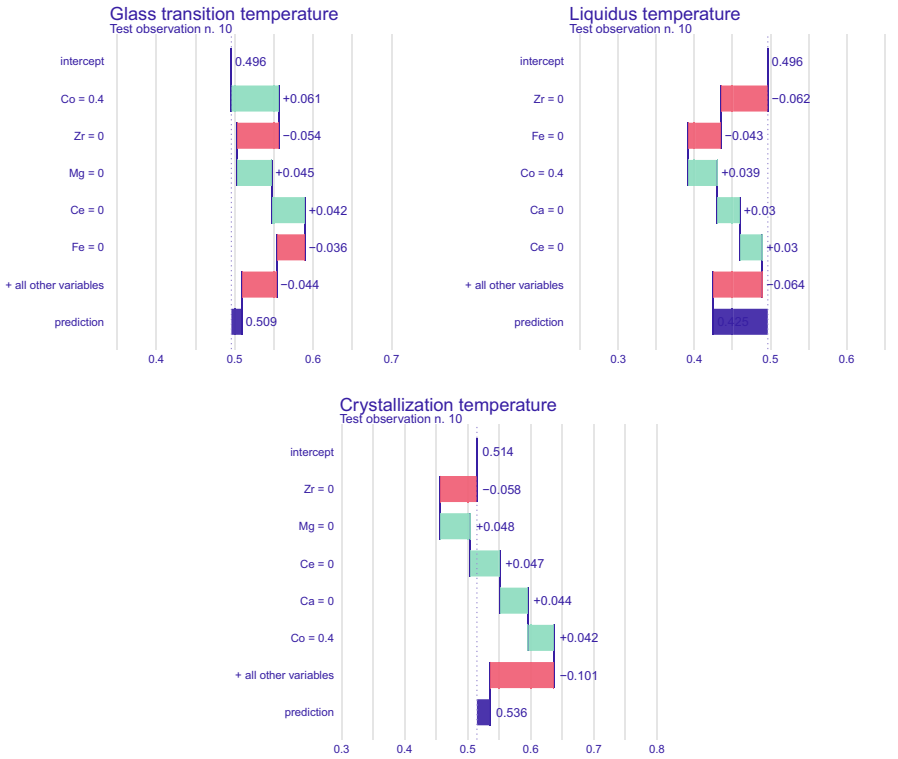

Fig. 6 SHAP break-down values for one observation obtained via ALPERC considering the training-test partition that leads to the best results at $iteration = 20$

to obtain a global view of the SHAP values for each predictor, considering all the test data. An example of these plots is provided for T_g, the Glass transition temperature, represented in Fig. 7, considering the 15 most relevant predictors identified in Fig. 5; if we consider the case of Copper (Cu) it is possible to observe that, if $Cu = 0.25$, T_g descreases by almost 0.025, while, if $Cu = 0.8$, the increase in T_g is equal to 0.025. So, using this partial dependence plot it is possible to observe how the various levels of some elements affect the different temperatures.

Moreover, from this plot it is possible to observe how these relationships are non-linear and also rather complex. A limitation of this kind of visualization is represented by the fact that these plots only provide information on the effect of the input variables taken individually. However it is possible to compute SHAP interaction values, which quantify the impact of the interactions after removing the impacts of the individual effects Molnar et al. (2020).

In Fig. 8, obtained using the treeshap R Package Komisarczyk et al. (2023), an example that displays a relevant interaction effect between Zirconium (Zr) and Copper (Cu) for the response T_l is represented.

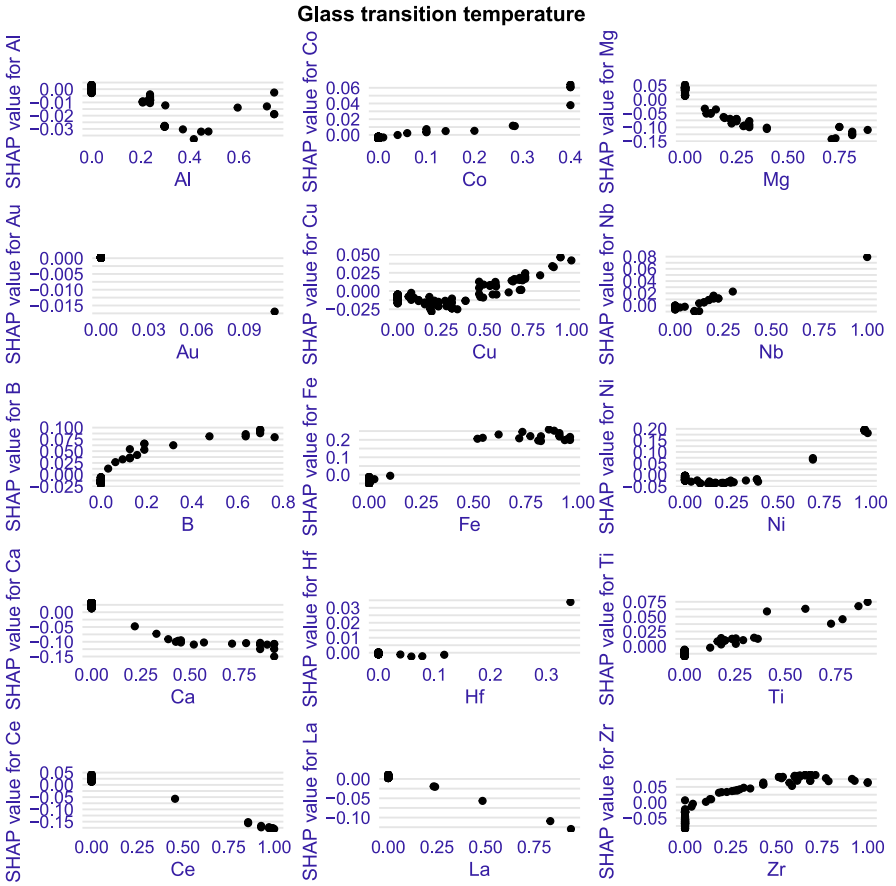

Fig. 7 Partial dependence plots on the test data for T_g considering the 15 most relevant predictors

In this example, it is important to point out that since the dataset is not a result of a designed experiment some confounding effects may arise while interpreting the interactions.

6 Results, interpretations and conclusions

In the first part of this paper, we provided a review of some contributions in the field of AL: firstly we reviewed a simulation study Arboretti et al. (2022b), in which the aim was to investigate which experimental design and ML model result most suitable for a joint application in physical experiments. From this simulation study, performed on 7 different test functions, it emerged that the best experimental design is MAXPRO_dis. The fact that this design is the best in the majority of situations may be explained by the fact that it is based on a space-filling criterion. This may favor flexible non-linear predictive models in capturing the behavior of the underlying function and, it

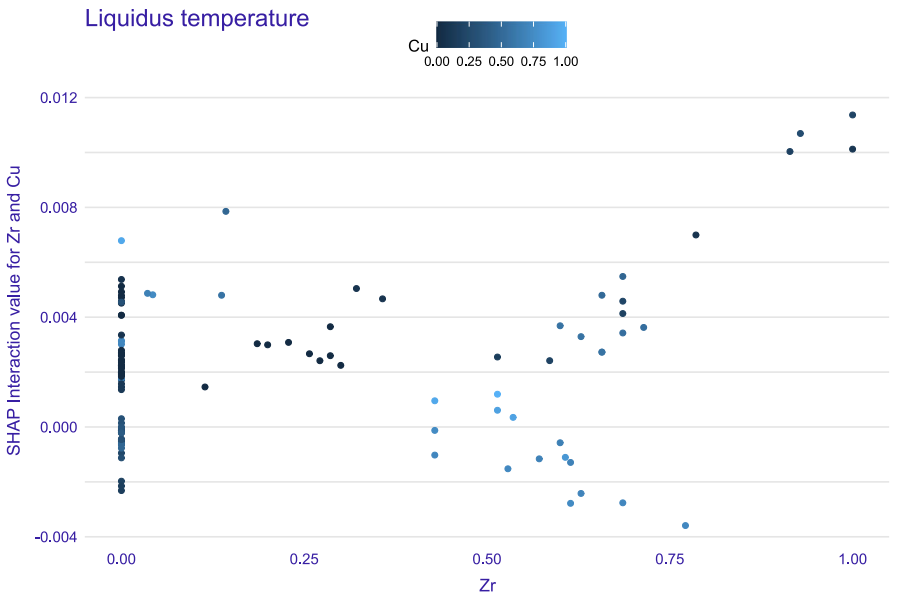

Fig. 8 Interaction plot of the variables Zr and Cu for the response T_l

just needs a limited number of levels, so it is also feasible for physical experiments. As regards the ML models, it emerged that the best choice is the Gaussian process, which resulted as the best choice in the vast majority of the different noise settings analyzed. This simulation study may represent the base for future research where the loss functions of the D-opt and I-Opt criteria are adapted to include heteroscedasticity. In Sect. 3 there is a review of ALPERC, a recently developed AL algorithm suitable for physical experiments when three or more responses are investigated. In Sect. 4 a simulation study compares ALPERC with other AL algorithms and also with a non-AL approach. From this review, it emerged that ALPERC provided a lower prediction error in comparison to the competitors, and also that AL algorithms performed better in almost all the analyzed situations, in contrast to a non-AL approach. Section 5 introduces the main novelty element of this paper, a case study, about amorphous metallic alloys, in which ALPERC is used together with an RF, in order to train and refine predictive models for emulating three different CTTs. In this case study, ALPERC proved to be more efficient than the non-AL strategy (Fig. 3) and this is a confirmation of the goodness of the algorithm. Moreover, also thanks to the adoption of the SHAP technique, the obtained model results could be easily interpreted by the analyst. To conclude, we can sum up the findings of this novel case study by saying that not only does ALPERC have a high potential for reducing predictive errors, but it also provides researchers with a more intuitive interpretation of the results.

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

Data availability The data used in the case study is taken from the work of Xiong et al. (2020). A R package including ALPERC functions is available at https://github.com/PegoraroL/ALPERC.

References

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46(1):243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Arboretti R, Bonnini S, Corain L, Salmaso L (2014) A permutation approach for ranking of multivariate populations. J Multivar Anal 132:39–57. https://doi.org/10.1016/j.jmva.2014.07.009

Arboretti R, Ceccato R, Pegoraro L, Salmaso L (2022) Active learning for noisy physical experiments with more than two responses. Chemom Intell Lab Syst 226:104595. https://doi.org/10.1016/j.chemolab.2022.104595

Arboretti R, Ceccato R, Pegoraro L, Salmaso L (2022) Design choice and machine learning model performances. Qual Reliabil Eng Int 38(7):3357–3378. https://doi.org/10.1002/qre.3123

Arboretti R, Ceccato R, Pegoraro L, Salmaso L (2022) Design of experiments and machine learning for product innovation: a systematic literature review. Qual Reliabil Eng Int 38(2):1131–1156. https://doi.org/10.1002/qre.3025

Arboretti R, Ceccato R, Pegoraro L, Salmaso L, Housmekerides C, Spadoni L, Pierangelo E, Quaggia S, Tveit C, Vianello S (2022) Machine learning and design of experiments with an application to product innovation in the chemical industry. J Appl Stat 49(10):2674–2699. https://doi.org/10.1080/02664763.2021.1907840

Binois M, Huang J, Gramacy RB, Ludkovski M (2019) Replication or exploration? sequential design for stochastic simulation experiments. Technometrics 61(1):7–23. https://doi.org/10.1080/00401706.2018.1469433

Bisgaard S (1992) Industrial use of statistically designed experiments: case study references and some historical anecdotes. Qual Eng 4(4):547–562. https://doi.org/10.1080/08982119208918936

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Chan K, Sort J (2015) Metallic glasses. Metals 5:2397–2400. https://doi.org/10.3390/met5042397

Freiesleben J, Keim J, Grutsch M (2020) Machine learning and design of experiments: alternative approaches or complementary methodologies for quality improvement? Qual Reliabil Eng Int 36(6):1837–1848. https://doi.org/10.1002/qre.2579

Gareth J, Daniela W, Trevor H, Robert T (2013) An introduction to statistical learning: with applications in R. Springer, Berlin

Gramacy RB, Lee HKH (2009) Adaptive design and analysis of supercomputer experiments. Technometrics 51(2):130–145. https://doi.org/10.1198/TECH.2009.0015

Jafary-Zadeh M, Praveen Kumar G, Branicio PS, Seifi M, Lewandowski JJ, Cui F (2018) A critical review on metallic glasses as structural materials for cardiovascular stent applications. J Funct Biomater 9(1):10019. https://doi.org/10.3390/jfb9010019

Joseph VR, Gul E, Ba S (2020) Designing computer experiments with multiple types of factors: the maxpro approach. J Qual Technol 52(4):343–354. https://doi.org/10.1080/00224065.2019.1611351

Khan MM, Nemati A, Rahman ZU, Shah UH, Asgar H, Haider W (2018) Recent advancements in bulk metallic glasses and their applications: a review. Crit Rev Solid State Mater Sci 43(3):233–268. https://doi.org/10.1080/10408436.2017.1358149

Komisarczyk K, Kozminski P, Maksymiuk S, Biecek P (2023) treeshap: fast SHAP values computation for tree ensemble models. r package version 0.1.1. https://github.com/ModelOriented/treeshap

LeDell E, Poirier S (2020) H2o automl: scalable automatic machine learning. Proc AutoML Workshop ICML 2020:1–16

Lujan-Moreno GA, Howard PR, Rojas OG, Montgomery DC (2018) Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. Expert Syst Appl 109:195–205. https://doi.org/10.1016/j.eswa.2018.05.024

Lundberg SM, Erion GG, Lee SI (2018) Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888

Molnar C, Casalicchio G, Bischl B (2020) Interpretable machine learning: a brief history, state-of-the-art and challenges. In: Koprinska I, Kamp M, Appice A, Loglisci C, Antonie L, Zimmermann A, Guidotti R, Özgöbek Ö, Ribeiro RP, Gavaldà R, Gama J, Adilova L, Krishnamurthy Y, Ferreira PM, Malerba D, Medeiros I, Ceci M, Manco G, Masciari E, Ras ZW, Christen P, Ntoutsi E, Schubert E, Zimek A, Monreale A, Biecek P, Rinzivillo S, Kille B, Lommatzsch A, Gulla JA (eds) ECML PKDD 2020 Workshops. Springer International Publishing, Cham, pp 417–431

Nair B, Priyadarshini BG (2016) Process, structure, property and applications of metallic glasses. AIMS Mater Sci 3(3):1022–1053. https://doi.org/10.3934/matersci.2016.3.1022

Olsson F (2009) A literature survey of active machine learning in the context of natural language processing. Tech. Rep. T2009:06, Swedish Institute of Computer Science. https://www.diva-portal.org/smash/get/diva2:1042586/FULLTEXT01.pdf

Shapley LS (2016) A value for n-person games. Princeton University Press, Princeton, pp 307–318

Sobol I (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 55(1):271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Staelin C (2003) Parameter selection for support vector machines

Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. J Mach Learn Res 15(48):1625–1651

Wei P, Lu Z, Song J (2015) Variable importance analysis: a comprehensive review. Reliabil Eng Syst Saf 142:399–432. https://doi.org/10.1016/j.ress.2015.05.018

Xiong J, Shi SQ, Zhang TY (2020) A machine-learning approach to predicting and understanding the properties of amorphous metallic alloys. Mater Des 187:108378. https://doi.org/10.1016/j.matdes.2019.108378

Yue X, Wen Y, Hunt JH, Shi J (2021) Active learning for gaussian process considering uncertainties with application to shape control of composite fuselage. IEEE Trans Autom Sci Eng 18(1):36–46. https://doi.org/10.1109/TASE.2020.2990401