Cheolhei Lee
Grado Department of Industrial and Systems Engineering,
Virginia Tech,
Blacksburg, VA 24061
e-mail: cheolheil@vt.edu

Kaiwen Wang
Department of Materials Science and Engineering,
Virginia Tech,
Blacksburg, VA 24061
e-mail: kaiwenwang@vt.edu

Jianguo Wu
Assistant Professor
Department of Industrial Engineering and Management,
Peking University,
Beijing 100080, China
e-mail: j.wu@pku.edu.cn

Wenjun Cai
Associate Professor
Department of Materials Science and Engineering,
Virginia Tech,
Blacksburg, VA 24061
e-mail: caiw@vt.edu

Xiaowei Yue[1]
Assistant Professor
Mem. ASME
Grado Department of Industrial and Systems Engineering,
Virginia Tech,
Blacksburg, VA 24061
e-mail: xwy@vt.edu

Partitioned Active Learning for Heterogeneous Systems

Active learning is a subfield of machine learning that focuses on improving the data collection efficiency in expensive-to-evaluate systems. Active learning-applied surrogate modeling facilitates cost-efficient analysis of demanding engineering systems, while the existence of heterogeneity in underlying systems may adversely affect the performance. In this article, we propose the partitioned active learning that quantifies informativeness of new design points by circumventing heterogeneity in systems. The proposed method partitions the design space based on heterogeneous features and searches for the next design point with two systematic steps. The global searching scheme accelerates exploration by identifying the most uncertain subregion, and the local searching utilizes circumscribed information induced by the local Gaussian process (GP). We also propose Cholesky update-driven numerical remedies for our active learning to address the computational complexity challenge. The proposed method consistently outperforms existing active learning methods in three real-world cases with better prediction and computation time.
[DOI: 10.1115/1.4056567]

Keywords: active learning, heterogeneous systems, partitioned Gaussian processes

1 Introduction

Active learning is a subfield of machine learning and artificial intelligence that maximizes information acquisition to train models' data efficiently. Contrary to passive learning such as Latin hypercube design (LHD) and factorial design [1], active learning sequentially selects design points in the modeling phase after observing intermediate models and outputs. It is also called query learning, sequential design, adaptive sampling, or optimal design in other literature, while they pursue the same objective: finding the best subset of inputs from the design space according to information criteria that evaluates the informativeness of input referring to uncertainty, disagreement, etc. Active learning has received increasing attention from various applications in which sampling is timely and costly demanding, such as quality engineering, response surface investigation, and image recognition [2,3].

Especially, active learning has been frequently utilized with Gaussian processes (GPs) for modeling of various systems spanning robotics, aerospace, and manufacturing processes [4,5] due to the capability of uncertainty quantification (UQ) and the simplicity [6,7]. However, many existing methods are confined to single GPs (SGPs) that impose inappropriate homogeneous information measures for systems with heterogeneity (e.g., discontinuity, and abrupt variations in gradient norms or frequencies), although heterogeneity is ubiquitous in engineering systems. For example, composite materials, one of the most versatile materials in various contemporary products, are anisotropic and highly nonlinear to external treatments [8], so they exhibit different behaviors in the design space [9]. Another example is the corrosion of alloys. Figure 1 illustrates the corrosive rates of aluminum alloys emulated with the finite element method (FEM) over two pairs of control variables. We can observe that the response surface shows spatial heterogeneity, so the design space can be partitioned into three subregions according to degrees of variation. In both cases, the efficiency of active learning with a homogeneous information criterion can be significantly deteriorated by the misleadingly measured information.

A straightforward way to address heterogeneity is to employ a locally adaptive information measure, which can be realized with so-called divide-and-conquer framework. For example, partitioned GPs (PGPs) overcome the limitations of SGPs in heterogeneous systems by allocating multiple independent local GPs on disjoint subregions. Subregions are defined or estimated according to distinguishable characteristics of target systems, so PGPs can efficiently accommodate heterogeneity. Moreover, the partitioning improves the scalability of GPs, one of their main drawbacks, by introducing sparsity in their covariance matrices. However, the partitioning methods in most of existing PGPs have the following drawbacks to directly use for active learning.

(1) Existing PGP methods [10,11] adopt Voronoi tessellation as a basis of partitioning, which is computationally prohibitive for high-dimensional design spaces.

[1]Corresponding author.
Contributed by the Computers and Information Division of ASME for publication in the Journal of Computing and Information Science in Engineering. Manuscript received July 24, 2022; final manuscript received December 16, 2022; published online January 10, 2023. Assoc. Editor: Guang Lin.

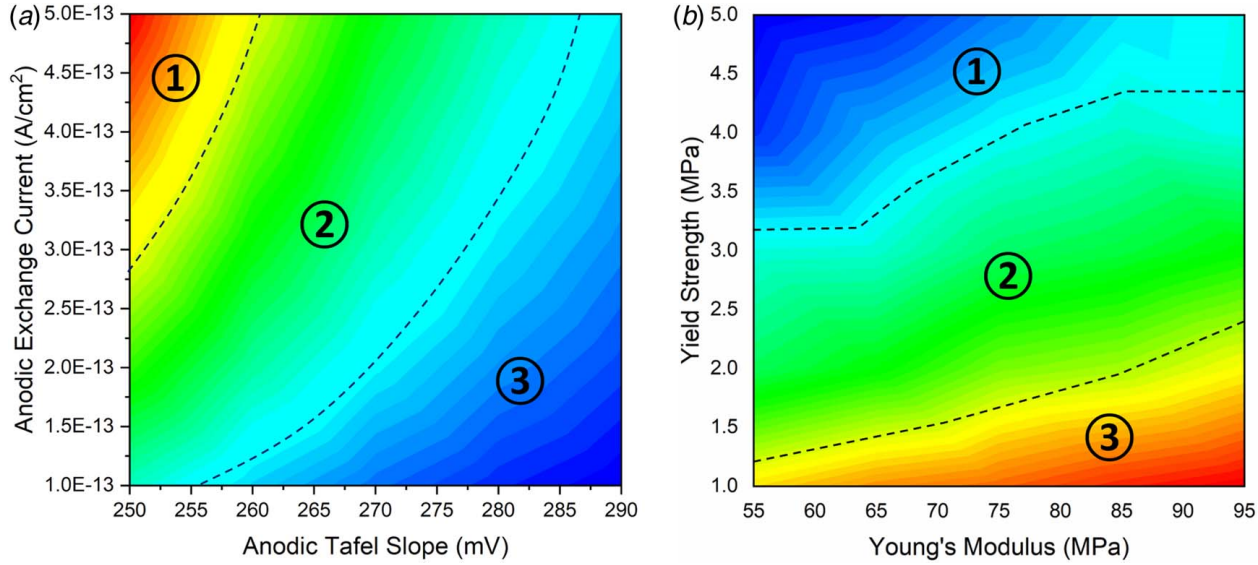

Fig. 1 Corrosive rates of alloys with different control variables: (*a*) Tafel slope and anodic exchange current and (*b*) Young's modulus and yield strength

(2) Many PGPs are established as hierarchical models in which partitioning models are submodels of local models [12,13]. Consequently, their partitions are driven by fitting their composite likelihoods, not directly referring to heterogeneity, which may yield implausible partitions (e.g., too many partitions, generation of trivial subregions).
(3) The hierarchical models using Markov chain Monte-Carlo (MCMC) methods are subject to computationally intractable posterior distribution, which should be utilized for active learning.

Moreover, existing works on PGPs utilize conventional active learning criteria [14,15], which are originally devised for a single model, so they are suboptimal for partitioned models in terms of learning and computational efficiency.

Motivated by the aforementioned limitations, we propose partitioned active learning for data acquisition in heterogeneous systems. To generate a locally adaptive measure, we divide the design space based on finite difference or output variance that are prevalent in many engineering systems. The proposed partitioning is based on mean-shift, so that partitioning can be efficiently adjusted with a single hyperparameter even in high-dimensional design spaces. Then, the localized integrated mean-squared error (IMSE) criterion is used, so that the IMSE criterion on each subregion cannot be distorted by other heterogeneous regions. To accelerate the searching time, we utilize the partitioned structure with global searching that seeks the most informative subregion and provide the Cholesky update method for computation of the local IMSE criterion. Our contributions in this article can be summarized as follows.

(1) A heterogeneity-based partitioning method is developed. The method outperforms the existing approaches in PGPs with (i) direct incorporation of heterogeneity features in partitioning; (ii) scalability to high-dimensional problems; and (iii) flexibility with a single tunable hyperparameter.
(2) A novel active learning strategy with a two-step searching that exploits the partitioned structure is established. The strategy first searches the region, which has the most potential to decrease the model variance, via global searching, and then the localized IMSE criterion is used to find the query location as local searching.
(3) Numerical remedies are provided to accelerate the proposed algorithm. Global searching is used to reduce the number of candidates by narrowing down the area for local searching, and the Cholesky factor update method is used to reduce the cost of local searching.

The proposed method is applied to prediction of three real-world cases: (i) fuselage deformation in aircraft manufacturing; (ii) tribo-corrosive rates of aluminum alloy; and (iii) inverse dynamics of seven-joint robot arm. In the case study, we observed that active learning with a homogeneous criterion can be even inferior to passive learning, which can be effectively resolved by the proposed method.

The remainder of this article is organized as follows. In Sec. 2, we review existing active learning methods for GPs and discuss several PGP modeling approaches. Section 3 elucidates the new partitioned active learning algorithm and provides applicable techniques for improving the learning efficiency and numerical costs. In Sec. 4, we implement our method on function approximation problems and apply to three real-world problems in Sec. 5 with existing learning algorithms. Finally, a brief summary of this article is provided in Sec. 6.

2 Literature Review

In this section, we review related literature of existing active learning strategies for GPs and their applications in real-world problems. Afterward, we discuss literature of PGP methods.

2.1 Active Learning for Gaussian Processes. Due to the capability of UQ in GPs, predictive uncertainty has been frequently involved in the construction of active learning criteria for GPs. Two criteria for GP regression, called active learning Mackay (ALM) and active learning Cohn (ALC), were compared in Ref. [16]. ALM refers to the variance as the information criterion that selects the most uncertain data.

Meanwhile, ALC refers to the IMSE criterion that seeks the point expected to reduce the model variance the most over the design space, and it has also been widely applied to different GP frameworks. The IMSE criterion was utilized for stochastic kriging, so as to balance between exploration over the design space and exploitation with additional replications [17]. Two other active learning algorithms were proposed for GP surrogate models of multimodal systems. The variance weighted active learning uses the weighted sum of variance criteria of modes, and the D-optimal weighted active learning uses the Fisher information matrix with the same engineering-driven weighting procedure [5]. By incorporating the

physics constraints, a failure-averse active learning algorithm has been proposed to realize efficient data collection as well as avoid failures [18]. Bayesian optimization (BO) is another interesting trend of sequential sampling that utilizes predictive uncertainty of GPs for global optimization problems, and it achieves outstanding performance [19,20]. Although BO and active learning show a subtle nuance in their objectives, they share the common principle and the sampling cost reduction. There are several criteria (also called acquisition functions) for BO, such as the expected improvement and the probability of improvement. They refer to the GP model of the objective function and choose the point that is the most likely to be optimal.

For strategies that do not take predictive uncertainty into account or less considered, space-filling designs were suggested to use in the Kernel Hilbert space induced by intermediate GP models [21]. The expected model output change was also used as one criterion for active learning [22]. The strategy chooses the location where is expected to induce the largest change in model outputs. The gradient of the GP model was involved in active learning so as to draw more samples from where the response abruptly changes [23,24]. Consequently, the gradient-based criterion focuses more on local variations, thereby reducing the prediction error more efficiently. Kim et al. [25] referred to the discrepancy of multiple GP models trained on different subsets of data. However, the aforementioned active learning strategies are mostly devised for SGPs that are inappropriate for modeling heterogeneous systems. Therefore, it is inefficient to directly apply them to PGPs since they cannot consider the partitioned structure. A comprehensive review of adaptive sampling strategies for GPs in the engineering domain can be found in Ref. [26].

2.2 Partitioned Gaussian Processes. Although there are techniques other than partitioning design spaces to overcome the limitation of stationary GP models such as input-dependent length scale, warping, and convolution kernels [24,27], we mainly focus on PGPs that explicitly partition the design space for multiple GP models. The piecewise GP was proposed using Voronoi tessellation with training dataset for partitioning the design space [10]. Advantages of Voronoi tessellation are simplicity, consistency, and distance-based algorithm that coincides with stationary GPs. They estimated the number of regions and centers with the Monte-Carlo approach and fitted independent GP models for subregions. Subsequently, the partitioning was generalized by merging convex Voronoi cells in order to generate nonconvex subregions and relaxed centers of cells [15]. The Treed GP was developed to use decision trees for partitioning [12]. They fitted independent local GPs on each leave associated with a subregion. Heaton et al. [11] proposed to partition the design space prior to GP modeling by hierarchical clustering referring to finite differences of samples. They insist that the approach allows avoiding the expensive MCMC algorithm in the aforementioned approaches. The mixture of GP experts proposed by Ref. [28] is another approach using multiple stationary GPs. Although components in their modeling process have own terminologies, the underlying idea is very close to the aforementioned methodologies.

For active learning in PGPs, an active learning algorithm was proposed, and it takes a point within boundaries and maximizes the space-filling property of design points [15]. However, it focuses more on detecting discontinuity in the design space rather than reducing the prediction error, although such discontinuity is rare in industrial and engineering applications. Active learning algorithms considering the posterior structure of partitioned design space were proposed [13,14]. However, their approaches are highly dependent on the tree classifier, while the tree partitioning mostly induces boundaries parallel to axes that may not be realistic in practice [11,15]. Moreover, the choice of design point candidates is dependent on the areas of partitioned regions, so it can be irrelevant when a plausible partition is not realizable with a few tree-partitioned regions.

3 Methodology

In this section, we propose the partitioned active learning algorithm. First, we briefly review a generic framework of partitioned modeling and two widely used active learning strategies for GPs. Then, we elucidate our strategy including partitioning based on heterogeneity and the partitioned information criterion. Finally, we provide applicable remedies to improve the computational cost of the proposed method.

As mathematical notations, we use lower letters for scalars and distinguish vectors with boldface. Upper letters indicate sets or matrices, and a set of indices is denoted as $[M] = \{1, \ldots, M\}$. In subscriptions, we parenthesize the number for the region index and use normal letters for indices of data. If the index of the subregion is apparent, we omit the region index and only use the index in the set for simplicity.

3.1 Partitioned Modeling for Heterogeneous Systems. A partitioned model of a heterogeneous system can be expressed as a function f defined over the design space $\Omega \subset \mathbb{R}^d$ mapping to $\mathbb{R}$. The partitioned model employs a region classifier $g : \Omega \to [M]$ that partitions Ω into M mutually disjoint subregions such that $\Omega = \bigcup_{m=1}^{M} \Omega_{(m)}$ in accordance with heterogeneity. Then, the partitioned model can be written as follows:

$$f(\mathbf{x}|g) = \sum_{m=1}^{M} 1_{\{g(\mathbf{x})=m\}} f_{(m)}(\mathbf{x}), \quad \mathbf{x} \in \Omega \tag{1}$$

where $1_{\{C\}}$ is an indicator function that has value 1 when C is true and 0 otherwise, and $f_{(m)}$ is a local GP assigned to $\Omega_{(m)}$. Although PGPs can take any valid kernel, we mainly consider stationary kernel family such as radial basis function (RBF) and Matérn. Suppose any $\mathbf{x}, \ \mathbf{x}' \in \Omega_{(m)}$ and $\mathbf{x} = [x_1 \ \ldots \ x_d]^\top$. The local GP defined over $\Omega_{(m)}$ with the RBF kernel is expressed as follows:

$$f_{(m)}(\mathbf{x}) \sim \mathcal{GP}\left(\mu_{(m)}(\mathbf{x}), \ k_{(m)}(\mathbf{x}, \mathbf{x}')\right)$$

$$k_{(m)}(\mathbf{x}, \mathbf{x}') = \tau_{(m)}^2 \prod_{j=1}^{d} \exp\left(-\frac{(x_j - x_j')^2}{l_{(m),j}^2}\right) + \sigma_{(m)}^2 1_{\{\mathbf{x}=\mathbf{x}'\}}$$

where $\mu_{(m)}(\mathbf{x})$ is the mean function assumed to be 0 without loss of generality, and $k_{(m)}$ is the kernel of which nonnegative $\tau_{(m)}^2$, $l_{(m),j}^2$, and $\sigma_{(m)}^2$ are referred as scale, length, and noise hyperparameters, respectively.

Suppose we have finite n observations on $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \Omega$ such that $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $y_i = f(\mathbf{x}_i) + \varepsilon_i$ of which $\varepsilon_i \sim \mathcal{N}(0, \ \sigma^2)$, and let Φ be the hyperparameter of the region classifier and Θ be that of local GPs. There are mainly two schemes in hyperparameter estimation of PGPs. One is to maximize the likelihood of the hierarchical structure of Eq. (1), and the other way is to estimate the region classifier in advance based on some criteria and fit local models by maximizing the likelihood with the fixed region classifier. We will discuss more about two schemes in Sec. 3.3.1. Suppose we have a region classifier from a prior or a separated algorithm, so that X and D can be partitioned as $X_{(m)}$ and $D_{(m)}$, respectively. Let the covariance matrix associated with $X_{(m)}$ be $K_{(m)}$ such that $(K_{(m)})_{ij} = k_{(m)}(\mathbf{x}_{(m),i}, \ \mathbf{x}_{(m),j})$ for $i, \ j \in [n_{(m)}]$, where $n_{(m)}$ is the number of samples in mth subregion. For the fixed g, the PGP model can be trained with D by maximizing the log marginal likelihood of local GPs, which is expressed as follows:

$$\ell(\Theta|g) \propto \sum_{m=1}^{M} \left(\mathbf{y}_{(m)}^\top K_{(m)}^{-1} \mathbf{y}_{(m)} - \log \det K_{(m)} - n_{(m)} \log 2\pi\right) \tag{2}$$

where $\mathbf{y}_{(m)} = [y_{(m),1} \ \ldots \ y_{(m),n_{(m)}}]^\top$. Note that the log likelihood in Eq. (2) is the sum of local GPs' due to their independence.

Consequently, possibly with some ordering process, the PGP produces a block diagonal covariance matrix, which implies that the numerical advantage of PGPs comes from the sparsity. Although the construction of entire covariance matrix is generally unnecessary in practice, it informs us that the model can be manipulated more efficiently by treating each local GP independently. Evidently, PGPs may exhibit discontinuity at partitioning boundaries, and it would be undesirable when the underlying truth is known to be continuous. Making PGPs continuity requires additional techniques, which is beyond the scope of this article, such as ensemble of the posterior local models [12] or patchwork kriging [29].

3.2 Review of Active Learning for Gaussian Processes. The essence of active learning is the information criterion, the function quantifying informativeness of unobserved data. By optimizing the information criterion in the design space, the learning machine determines the design point to learn and the queries to the oracle. The variance and the IMSE criteria are widely considered in active learning for GPs due to their versatility and simplicity [16,14,5], so we briefly review two criteria following terminologies in Ref. [16] and then establish our new criterion.

Suppose we intend to determine the next sampling location $\mathbf{x}_{n+1} \in \Omega$ with the GP without partitioning. The variance criterion is expressed as follows:

$$J_{\mathrm{V}}(\mathbf{x}) = k(\mathbf{x}) - \mathbf{k}(\mathbf{x},\ X)K^{-1}\mathbf{k}(\mathbf{x},\ X)^{\top} \tag{3}$$

where f is the posterior GP given X. ALM maximizes the variance criterion so as to select the location with the greatest predictive variance. Meanwhile, the IMSE criterion is expressed as follows:

$$\begin{aligned} J_{\mathrm{IMSE}}(\mathbf{x}) &= \int_{\Omega} \mathrm{Var}\big(\, f(\mathbf{s}|\,\mathbf{x})\big)\, p(\mathbf{s})\, \mathrm{d}\mathbf{s} \\ \mathrm{Var}\big(\, f(\mathbf{s}|\,\mathbf{x})\big) &= k(\mathbf{s}) - \mathbf{k}(\mathbf{s},\ X_{n+1})K_{n+1}^{-1}\,\mathbf{k}(\mathbf{s},\ X_{n+1})^{\top} \\ X_{n+1} &= [X_n \quad \mathbf{x}]^{\top} \end{aligned} \tag{4}$$

where $\mathbf{s} \in \Omega$ with the density (or importance) $p(\mathbf{s})$ and K_{n+1} is the covariance matrix associated with X_{n+1}. Minimizing the IMSE criterion selects the location, which is expected to reduce predictive uncertainty the most over Ω, and we refer to the active learning with the IMSE criterion as ALC.

There are more behind derivations of both criteria, while we mention shortly herein. It turns out that ALM is equivalent to the maximum entropy design since the choice leads to maximizing the determinant of covariance matrix. Meanwhile, ALC can be explained by minimizing the generalized mean-squared error (MSE) in statistical learning, which can be decomposed into the bias and the variance. Although the variance criterion is straightforward and numerically inexpensive, ALC empirically has shown better performance than ALM [14,16]. Moreover, ALC can comprehensively consider the importance of $\mathbf{s} \in \Omega$ in the information criterion. It allows us to incorporate prior knowledge and to give more weight on a specific region, thereby making the algorithm more distinguishable from the space-filling design.

3.3 Partitioned Active Learning Strategy. Dividing the design space in heterogeneous systems allows flexible modeling for GPs, while most of existing partitioning methods do not refer to heterogeneous features therein. Moreover, the most widely used active learning strategies in the previous section also do not take account partitions in their information criteria. This section illustrates our proposed active learning strategy that includes a separated partitioning scheme based on heterogeneity and the new information criterion that exploits the partitioned structure.

3.3.1 Partitioning Based on Heterogeneity. The main objective of partitioning in our method is to exclude the adverse effect from the heterogeneous subsystems. There are two ways to determine partitions in PGPs: model driven and model free. Model-driven partitioning establishes the region classifier by maximizing the likelihood of the hierarchical model in Eq. (1). It is promising to exploit the training data to construct a well-fitted model, while it may induce partitions that are difficult to interpret. For example, the number of subregions can be unnecessarily high, which leads to the loss of correlation information with overfitting. Another drawback is the expensive computational cost from partitioning and MCMC algorithms. Existing methods utilize Voronoi tessellation or decision tree as a basis. The numerical complexity of Voronoi tessellation is known to be $\mathcal{O}(n \log n + n^{\lceil d/2 \rceil})$ [30], which exponentially increases with the dimension, so it can be numerically problematic in high-dimensional problems such as our case study in Sec. 5.3. Furthermore, it is inapplicable when the number of samples is less than the problem dimension. Decision tree is also dependent to the input dimension in its complexity, which is $\mathcal{O}(dn \log n)$.

Meanwhile, model-free approach builds the region classifier separately before fitting local models. Although it may induce an inferior fit of the partitioned model with respect to the training data, the region classifier can be directly established based on predetermined criteria (e.g., heterogeneous factors). Moreover, the partitioning can be examined and modified with background knowledge by experts before plugging the classifier into the partitioned model. Therefore, if we have a prior knowledge in partitioning (e.g., the number of subregions, features that determines the heterogeneity), the model-free approach is more reasonable. Agglomerative clustering in Ref. [11] is model free, and it utilizes the finite difference in observed data as the heterogeneous feature. However, it is also based on Voronoi tessellation and may yield undesirable singleton clusters, in which local GPs cannot be trained, as shown in Fig. 2. Although it is possible to employ the agglomerative clustering for some cases, we provide another path of partitioning, which tackles the aforementioned limitations.

To divide the design space based on heterogeneity, we first need to define heterogeneous characteristics. There could be many candidates that induces heterogeneity in systems, and the example includes variations and variances. Let h be the heterogeneous feature. The degree of variation at a point can be quantified with the gradient norm, while the best accessible reference with finite observations would be finite differences between the point and neighbors, which is

$$h(\mathbf{x}) = \mathrm{avg}\left(\frac{\mathrm{d}(y_i,\ y)}{\mathrm{d}(\mathbf{z}_i,\ \mathbf{x})} \,\middle|\, \mathbf{z}_i \in N_r(\mathbf{x}) \right) \tag{5}$$

where $N_r(\mathbf{x}) = \{\mathbf{z} | d(\mathbf{x},\ \mathbf{z}) \le r,\ \mathbf{z} \in \Omega\}$. The radius r should be large enough to cover more than one adjacent point, but not too large to exclude irrelevant points. A plausible value is 1–1.5 times of the minimum distance in observations. The main difference between Eq. (5) and the dissimilarity in [11] is that Eq. (5) considers all adjacent samples for every design point, while the other calculates the dissimilarity of every pairs like a graph model to conduct agglomerative clustering. For heteroscedastic stochastic systems, variance can be a good reference. The regional variance can be approximated as follows:

$$h(\mathbf{x}) = \mathrm{Var}\big(y_i |\, \mathbf{x}_i \in N_r(\mathbf{x})\big) \tag{6}$$

Once heterogeneous features are evaluated for observed data, the kernel density estimation can be used to cluster them. In this article, we utilize mean-shift [31] to detect modes of heterogeneity in the design space as follows. There are several advantages in the use of mean-shift for clustering. First, it does not require a specific number of clusters to implement the procedure. There are many ways to estimate the bandwidth with data a prior [32], and it can be easily modified with a single bandwidth parameter for a specific

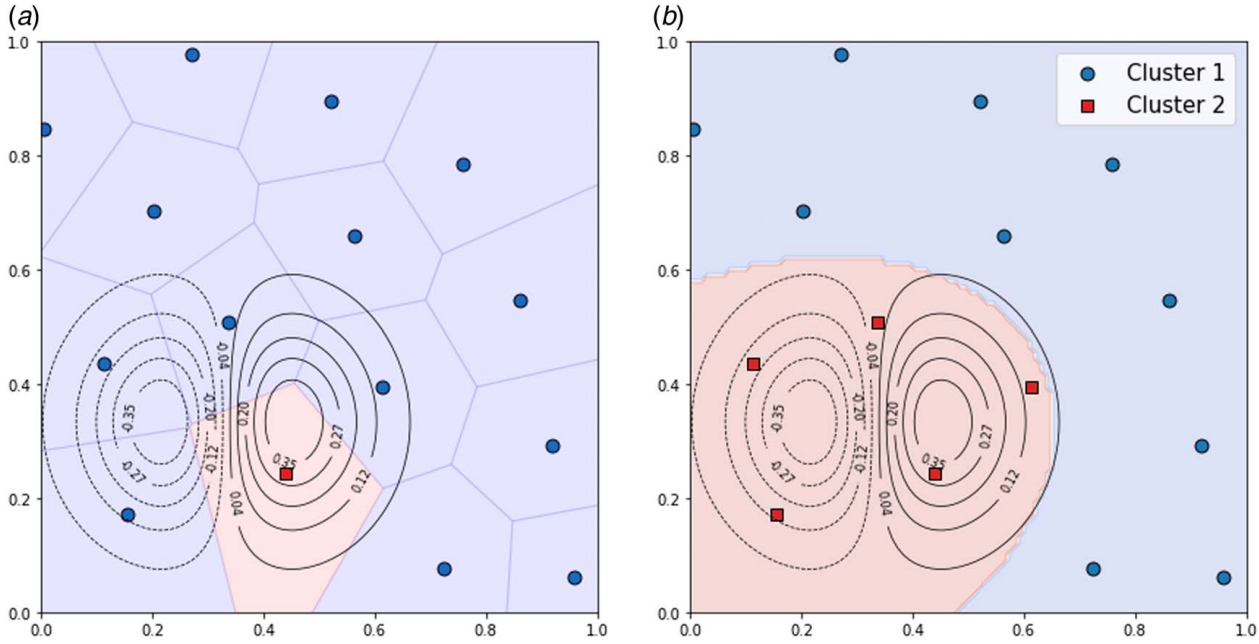

Fig. 2 Partitioning in 2D simulation study: (*a*) agglomerative clustering [11] and (*b*) SVC on mean-shift

number of clusters. Second, its complexity is independent to input dimension, so it is numerically more advantageous than the tessellation-based methods in high-dimensional problems. The training complexity is the squared number of data, but can be reduced to $\mathcal{O}(n \log n)$ with additional algorithms. Let $\tilde{\mathbf{x}}_i$ be the augmented vector of $\mathbf{x}_i$ with $h_i \equiv h(\mathbf{x}_i)$ and consider a nonnegative bandwidth parameter γ for the kernel to climb the density of the heterogeneity feature (e.g., finite difference). Then, we can find the heterogeneity mode associated with $\tilde{\mathbf{x}} \in \Omega \times \mathbb{R}$ by iteratively updating the mean-shift vector ($\tilde{\mathbf{z}}$) as follows:

$$\tilde{\mathbf{z}}_{j+1} = \tilde{\mathbf{z}}_j + \frac{\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \exp\left(-(1/2)\left\|(\tilde{\mathbf{z}}_j - \tilde{\mathbf{x}}_i)/\gamma\right\|^2\right)}{\sum_{i=1}^{n} \exp\left(-(1/2)\left\|(\tilde{\mathbf{z}}_j - \mathbf{x}_i)/\gamma\right\|^2\right)} \tag{7}$$

for $j = 1, 2, \ldots$, until its convergence. The sequence of Eq. (7) begins from each of $\tilde{\mathbf{z}}_1 \in \{\tilde{\mathbf{x}}_i\}_{i=1}^n$, and the number of converging modes is the number of clusters. The number of partitions is important, while the mean-shift procedure may not provide a desirable partition. Especially, in the case of no prior knowledge about the presence of heterogeneity, decision of the number of partitions can be very challenging. In this case, the number of clusters should be set in a conservative manner (i.e., as small as possible) to keep the possible correlation between subregions, and the mean-shift result should be examined and modified according to some criteria (e.g., the size of each cluster, the number of clusters).

However, the resulting mean-shift model cannot be used for active learning, since it requires the heterogeneous feature of unobserved data. Thus, we need to employ a discriminative function that classifies input to the heterogeneity classes induced by the mean-shift procedure. Although there are several candidate classifiers (e.g., support vector classification (SVC), decision tree, k-nearest neighbor) for the discriminative function, nonlinear SVC is mainly employed in this article owing to its flexibility and effectiveness in high-dimensional problems. The discriminative function will be the region classifier, and the procedure is provided in Algorithm 1.

3.3.2 Partitioned Information Criterion. In this section, we dedicate to the construction of the information criterion induced by the heterogeneity-based region classifier. When the IMSE criterion is considered for a candidate location $\mathbf{x} \in \Omega_{(m)}$ with PGPs, it can be written as follows:

$$J(\mathbf{x}) = \sum_{l \neq m} \int_{\Omega_{(l)}} \mathrm{Var}\left(f_{(l)}(\mathbf{s}_{(l)})\right) p(\mathbf{s}_{(l)})\, \mathrm{d}\mathbf{s}_{(l)} + \int_{\Omega_{(m)}} \mathrm{Var}\left(f_{(m)}\mathbf{s}_{(m)}|\mathbf{x}\right) p(\mathbf{s}_{(m)})\, \mathrm{d}\mathbf{s}_{(m)} \tag{8}$$

where $\mathbf{s}_{(m)} \in \Omega_{(m)}$. The interpretation of each term in Eq. (8) is quite worthwhile. The first term is the sum of IMSEs except for $f_{(m)}$, and it is invariant to the choice of $\mathbf{x} \in \Omega_{(m)}$. The second term is equivalent to Eq. (4), in which $f_{(m)}$ is only considered, so that Eq. (8) will only take account of the local region where the candidate is located. Briefly, there are two main differences between (4) and (8): (i) consideration of IMSEs over other local regions; and (ii) the localized IMSE criterion. We focus on each term subsequently considering their meanings, thereby efficiently minimizing Eq. (8).

Heuristically, Eq. (8) is more likely to be minimized when the most uncertain local GP, which has more potential to be reduced with additional observations, is considered in the second term, since the local GP. Each IMSE in the first term indicates the regional uncertainty of PGP, so it can be used for investigating the most uncertain region. Let us denote the regional uncertainty of each local GP as $\mathbb{V}_{(m)}$ for $m \in [M]$. The global searching choose the most uncertain region from the following categorical distribution:

$$m^* \sim \mathrm{Cat}\left(\frac{\mathbb{V}_{(1)}}{\sum_{m=1}^{M} \mathbb{V}_{(m)}}, \ldots, \frac{\mathbb{V}_{(M)}}{\sum_{m=1}^{M} \mathbb{V}_{(m)}}\right) \tag{9}$$

A straightforward way to choose the most uncertain region is to select the maximum in the sequence of regional uncertainty, while sampling from Eq. (9) can prevent some undesirable states such as falling into a specific subregion due to insufficient information about other subregions, or presence of multiple subregions with comparable variance.

Once the most uncertain region is determined by Eq. (9), we focus on the second term within $\Omega_{(m^*)}$ as the local searching with

the following criterion:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\Omega_{(m^*)}} \int_{\Omega_{(m^*)}} \mathrm{Var}\left(f_{(m^*)}(\mathbf{s}_{(m^*)}|\,\mathbf{x})\right) \times p(\mathbf{s}_{(m^*)})\,\mathrm{d}\mathbf{s}_{(m^*)} \tag{10}$$

Since the local GP in Eq. (10) reflects the local behavior of underlying function excluding heterogeneity from other regions, it can lead to improvement in the exploitation of active learning by avoiding implausible predictive uncertainty. We call the sequential criteria (9) and (10) by partitioned IMSE (PIMSE) and the active learning with PIMSE as partitioned ALC (PALC). The PIMSE criterion asymptotically converges to a steady state as the number of observations increases as the following proposition.

PROPOSITION 1 (Convergence of PALC). *The PIMSE criterion uniformly converges on* Ω *as* $n \to \infty$.

Proof. Let us denote the local IMSE criterion in Eq. (10) with n observations as J_n. The sequence of $\{J_n,\ J_{n+1}, \ldots\}$ is monotonically decreasing for all $\mathbf{x} \in \Omega$ by Theorem 2 of Ref. [17] and clearly lower bounded by zero. Thus, it converges. ∎

It is noteworthy that the uniform convergence in Proposition 1 implies that it leads to the best estimator of f, which of uncertainty is irreducible. That is, if one employed the stochastic local kriging, the PIMSE criterion will be asymptotically dominated by the intrinsic uncertainty of the nugget effect.

Algorithm 1 Partitioned ALC

Prerequisite: $D = (X,\ \mathrm{y})$, N_{iter}, N_{ref}, N_{cand}
Partitioning based on Heterogeneity
  Calculate h with D
  Implement mean-shift (7) on $(X,\ \mathrm{h})$ to generate M clusters
  Train g on X and the generated cluster labels m
Train f on D with g by maximizing (2)
while $i < N_{\mathrm{iter}}$ **do**
  Global Searching:
    Generate $X_{\mathrm{ref}} \subset \Omega$ with N_{ref} space-filling design
    Calculate $\mathbb{V}_{(m)}$ for $m \in [M]$ with X_{ref} with (11)

$$m^* \sim \mathrm{Cat}\left(\frac{\mathbb{V}_{(1)}}{\sum_m \mathbb{V}_{(m)}}, \ldots, \frac{\mathbb{V}_{(M)}}{\sum_m \mathbb{V}_{(m)}}\right)$$

  Local Searching:
    Generate $X_{\mathrm{cand}} \subset \Omega_{(m^*)}$ with N_{cand} space-filling design
    Obtain x^* by solving (10) with $\Omega_{(m)} = X_{\mathrm{cand}}$
  Obtain y^* at x^*
  $D = D \cup \{(\mathrm{x}^*, y^*)\}$
  Update g and f on D
  i = i+1
end while

The pseudocode of PALC is provided in Algorithm 1. The prerequisite values $\{N_{\mathrm{iter}}, N_{\mathrm{ref}}, N_{\mathrm{cand}}\}$ stand for the number of attainable samples (i.e., budget), reference design points for Eq. (9) and candidate points for Eq. (10), and they are required for practical implementation of the proposed algorithm. X_{ref} is a reference set composed of N_{ref} space-filling points ($\{\mathbf{s}_i\}_{i=1}^{N_{\mathrm{ref}}} \subset \Omega$), which is required to implement the global searching as follows:

$$\mathbb{V}_{(m)} \approx \frac{1}{N_{ref}} \sum_{i=1}^{N_{ref}} \mathrm{Var}\left(f_{(m)}(\mathbf{s_i}|\mathbf{x})\right) \tilde{p}(\mathbf{s}_{(m),i}) \tag{11}$$

where $\tilde{p}(\mathbf{s}_{(m),i})$ is the approximated probability mass function at $\mathbf{s}_{(m),i}$. Then, N_{cand} is passed to generate the candidate pool $X_{\mathrm{cand}} \subset \Omega_{(m^*)}$, and the subset of the reference set associated with the chosen subregion will be used for solving Eq. (10). In this article, we have generated X_{ref} and X_{cand} with new LHD in every step in order to encourage exploration.

Aside from the main algorithm, there is no universal concrete theorem for the optimal portion of initial sampling, while some empirical suggestions can be found in Refs. [5,26]. However, we can conjecture that the number of initial samples has a trade-off property. If the initial sampling is weighed too much, the advantage of active learning will be diluted. Otherwise, active learning can be hindered by low-quality information from unreliable intermediate models. Also, in the case of the region classifier g, the number should be enough to obtain an acceptable g. For termination of the active learning procedure, early stopping can be a reasonable choice other than the sampling budget in Algorithm 1 when a separated testing dataset or cross-validation scheme is available.

3.4 Cholesky Update-Based Numerical Remedies to Tackle the Computational Complexity Challenge. The IMSE criterion is numerically more demanding than the variance criterion, since it involves the inversion of K_{n+1}, which should be updated for every candidate. That is, calculation of the IMSE criterion requires $\mathcal{O}(n^3)$ for each candidate. Moreover, when N candidates are provided to the active learning module, the computational cost is multiplied by the number. Although the significance of their effects may vary with situations, the effect of candidate number can be more considerable than the inversion cost. Therefore, in order to improve the numerical aspect of PALC, we should provide some remedies for both matrix inversion and the number of candidates.

The global searching reduces the number of candidates by narrowing down the region of interest. Generally, candidates for active learning are given with space-filling or dense-grid over the design space. Therefore, taking the subset of candidates in the most uncertain region with the global searching leads to reduction in the number of matrix inversions proportional to the ratio of the chosen region from Eq. (9).

The matrix inversion cost of PGP is automatically alleviated by partitioning the design space with the block diagonal covariance matrix. That is, the inversion cost reduces from $\mathcal{O}(n^3)$ to at most $\mathcal{O}({n_{(m)}}^3)$, where $n_{(m)} < n$. Another applicable remedy is updating the inverse of $K_{n_{(m)}+1}$ in Eq. (10) exploiting $K_{n_{(m)}}^{-1}$ iteratively. Although it is possible to apply the Sherman–Morrison formula to get the updated inverse matrix [14], the Cholesky decomposition [6] for solving the linear system $K_{n_{(m)}}^{-1}\mathbf{k}$ could be more preferable considering the numerical stability and cost. Given that the Cholesky factor (a lower triangular matrix) of $K_{n_{(m)}}$ is known, updating the Cholesky factor of $K_{n_{(m)}+1}$ only requires the forward substitution step of the size $n_{(m)}$ triangular system; thus, it needs only $\mathcal{O}(n_{(m)}^2)$ instead of $\mathcal{O}({n_{(m)}}^3)$. A more detailed procedure of the Cholesky update is provided as follows.

Suppose we have the Cholesky factor L of K_n, which is the covariance matrix of X_n, such that $K_n = LL^\top$. We aim to get the Cholesky factor $\hat{L}$ of

$$K_{n+1} = \begin{bmatrix} K_n & \mathbf{k}_{\mathbf{n}}^* \\ \mathbf{k}_{\mathbf{n}}^{*\top} & k(\mathbf{x}^*) \end{bmatrix}$$

where $\mathbf{x}^*$ is a candidate input and $\mathbf{k}_n^* = k(X_n,\ \mathbf{x}^*)$. Since K_{n+1} shares the same part of K_n, it turns out that the first $n \times n$ elements of $\hat{L}$ is equivalent to L. Therefore, we can apply the Cholesky–Banachiewicz algorithm for $\hat{L}$ as follows:

$$\hat{L}_i = \begin{cases} L_{i,i}^{-1}\left(\mathbf{k}_{\mathbf{n}i}^* - \sum_{j=1}^{i-1} \hat{L}_j L_{i,j}\right), & i \in [n] \\ \sqrt{k(\mathbf{x}^*) - \sum_{j=1}^{n} \hat{L}_j}, & i = n+1 \end{cases}$$

Rather than calculating the PIMSE directly, PALC can be faster by skipping the redundant computation by applying the Cholesky updating approach. The Cholesky factor L can be used for the predictive variance of GP in Eq. (3) for $\mathbf{s} \in X_{\mathrm{Ref}}$, and the global searching criterion of Eq. (9) can be expressed as follows:

$$\mathrm{Var}_n^2(\mathbf{s}) = k(\mathbf{s}) - \mathbf{v}_n{}^\top \mathbf{v}_n, \quad \mathbf{v}_n = L^{-1}\mathbf{k}_n$$

where $\mathbf{k}_n = k(X_n,\ \mathbf{s})$. In the same manner, the optimal solution of local searching in Eq. (10) can be obtained by minimizing

$$\mathrm{Var}^2_{(m^*),n_{(m^*)}+1}(\mathbf{s}) = k(\mathbf{s}) - \mathbf{v}^{*\top}\mathbf{v}^* \tag{12}$$

$$\mathbf{v}^* = \hat{L}^{-1}k(X_{n+1},\ \mathbf{s}) \tag{13}$$

Since we already have the solution of (13) partially with $\mathbf{v}$ (i.e., $\mathbf{v}^*[1:n] \equiv \mathbf{v}_n$), we need only $v^*_{n+1} := \mathbf{v}^*[n+1]$, which can be calculated with a forward substitution as follows:

$$v^*_{n+1} = k(\mathbf{x}^*, \mathbf{s}) - \sum_{j=1}^{n} \hat{L}_k \mathbf{v}_j$$

Since $k(\mathbf{s})$ in Eq. (12) is invariant to $\mathbf{s}$ and $\mathbf{x}^*$, only the second term $\mathbf{v}^{*\top}\mathbf{v}^*$ is usually considered as the simplified PIMSE criterion, which must be maximized in PALC.

4 Simulation Study

In this section, we evaluate our active learning algorithm with simulation data. Two functions are considered that can be visualized straightforwardly. Both functions contain heterogeneous response surfaces, and observation noise is involved. In order to reduce the variability from the random initial dataset, each simulation is replicated ten times. As our benchmark methods, ALC and ALM are considered for SGPs, and the following partitioned active learning methods are considered: (i) the variance criterion for the local searching (PALM); (ii) PALC without global searching (PALC-NoG). In addition, uniform random sampling (Rand) and LHD are also considered as passive learning. For evaluation of models, one thousand space-filling design points for each design space are used with root-mean-squared error (RMSE) as the metric. Total computational times spent on querying are also compared excluding the model training time.

4.1 One-Dimensional Data. We apply our proposed active learning algorithm to a one-dimensional simulation function:

$$f(x) = 2x\sin(8\pi x^3)$$

which is defined on [0, 1] as the dotted line in Fig. 3. Zero-mean Gaussian noise is imposed with variance $\sigma^2 = 10^{-4}$. We allocate ten samples for initial training using Maximin LHD, and 20 samples are sequentially obtained via active learning. The function is differentiable, with heterogeneous frequency and amplitude over the domain. With the variation feature in Eq. (5), heterogeneity-based partitioning provided two to three partitions, and the number is set to two considering the number of initial samples and the dimension by adjusting the bandwidth parameter with grid searching over $[10^{-5}, 10^3]$. Afterward, logistic regression is used for the region classifier, which resulted in the decision boundary around $x = 0.38$–0.65.

Figure 3 shows each GP model fitted with initial samples. First, Fig. 3(*b*) illustrates that the PGP prevents misled active learning by providing appropriate predictive uncertainty. Also, we can observe from Fig. 3(*a*) that the next design point to be queried in the SGP model is chosen from the low-frequency region with both variance and IMSE criteria. Meanwhile, in the PGP model, the variance criterion takes the point from the boundary, while the IMSE does not. It implies that the IMSE criterion can be more promising when the adjacent local GPs show comparable predictive uncertainties.

Table 1 summarizes the results after the full data acquisition, where the numbers in parentheses indicate standard deviation of the results from replications. We grouped the considered methods into three: (i) passive learning (Rand; and LHD); (ii) variance methods (ALM; and PALM); and (iii) IMSE-based methods (ALC; PALC-NoG; and PALC). We can observe that the predictive accuracy of PALC outperforms the others. In the computational time, variance methods (ALM and PALM) are definitely faster than the IMSE-based methods, while they are deficient in predictive accuracy; ALM even worse than the random sampling and the LHD. Among the IMSE-based methods, PALC is faster than the others. Moreover, if we focus on PALC and PALC-NoG, we can observe that the global searching does not only reduce the computational time but also improve the learning efficiency.

4.2 Two-Dimensional Data. We extend the simulation study to a two-dimensional function (shown in the left of Fig. 4(*a*)), which is also used in Refs. [13,14]. The function is composed of two regions: even and uneven, and zero-mean Gaussian noise with variance $\sigma^2 = 10^{-6}$ is imposed as the observation noise. In a similar manner, we begin with 15 samples with LHD and obtain 15 additional samples via active learning.

For partitioning the design space, we used finite differences as the heterogeneous feature and used SVC after labeling initial samples based on the mean-shift result as shown in Fig. 4(*a*). Some initial dataset yielded three partitions with affordable evenness, while we adhered to two partitions due to the majority of two partitions in all replications. Since region 1 is less interesting than region 2, we can observe that the PIMSE criterion induced by two independent local GPs provides more relevant information of design points as shown in Fig. 4(*c*). Consequently, the IMSE criterion with the SGP fails to pick from the more interesting region due to the misled information criterion. The results with standard deviation (in parentheses) are summarized in Table 2, and we can see that PALC surpasses the other methods again in both predictive accuracy and computational time among the IMSE-based methods.

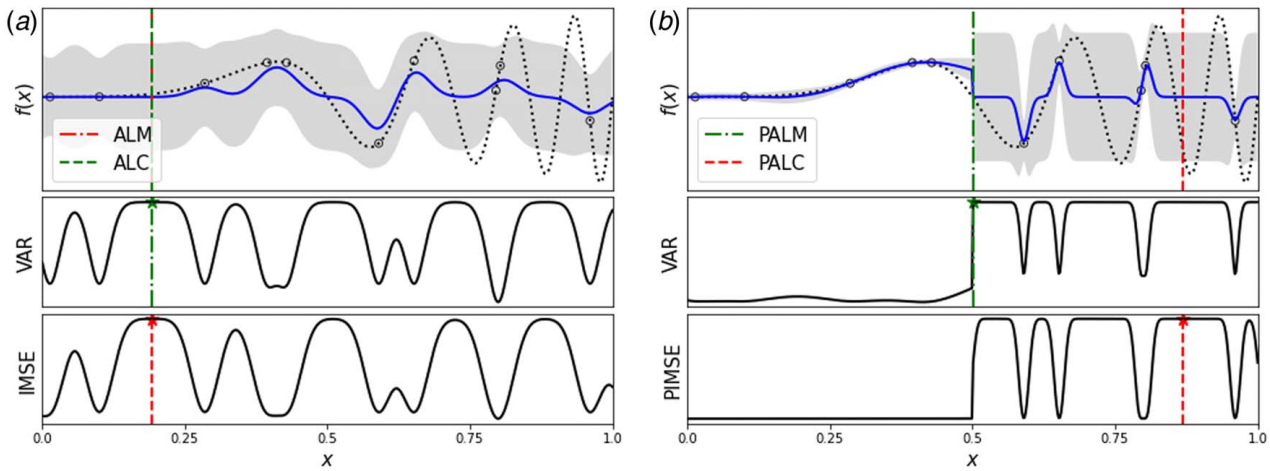

Fig. 3 Results of different active learning algorithms in 1D simulation: (*a*) SGP and (*b*) PGP with two local GPs

Table 1 One-dimensional simulation study results

Methods	RMSE	Time (s)
Rand	0.395 (0.152)	—
LHD	0.332 (0.133)	—
ALM	0.352 (0.159)	0.341 (0.140)
PALM	0.072 (0.270)	0.144 (0.049)
ALC	0.272 (0.205)	11.468 (0.424)
PALC-NoG	0.051 (0.277)	7.049 (0.024)
PALC	**0.048 (0.273)**	**3.553 (0.137)**

Note: Predictive accuracy of PALC outperforms the others, which is denoted in bold.

5 Case Study

In this section, we apply our approach to construct surrogate models for three different real-world cases. The purpose of the surrogate models is to embed them into automated systems and to provide UQ in posterior analysis. Apart from the benchmark methods in the simulation study, we also considered agglomerative clustering in Ref. [11] for case study, while it was inapplicable due to generation of singleton clusters as in Sec. 3.3.1. For mean-shift in our approach, we adjusted the number of partitions referring to evenness of clusters for all cases as the simulation study, since we have no prior knowledge about underlying partitioning. Case studies include higher input dimensions than the previous simulations.

5.1 Residual Stress of Composite Fuselages in Aerospace Manufacturing. We apply our proposed active learning strategy to construct the predictive model of residual stress in the composite fuselage assembly process. In the aircraft manufacturing process, composite fuselages are built in several subsections independently, so they are subject to the discrepancy in the junction part. The composite fuselage is reshaped using multiple fixed actuators in the automatic shape control. In order to achieve the optimal manufacturing process, the shape control needs to consider not only the deformation but also the residual stress in the structure due to

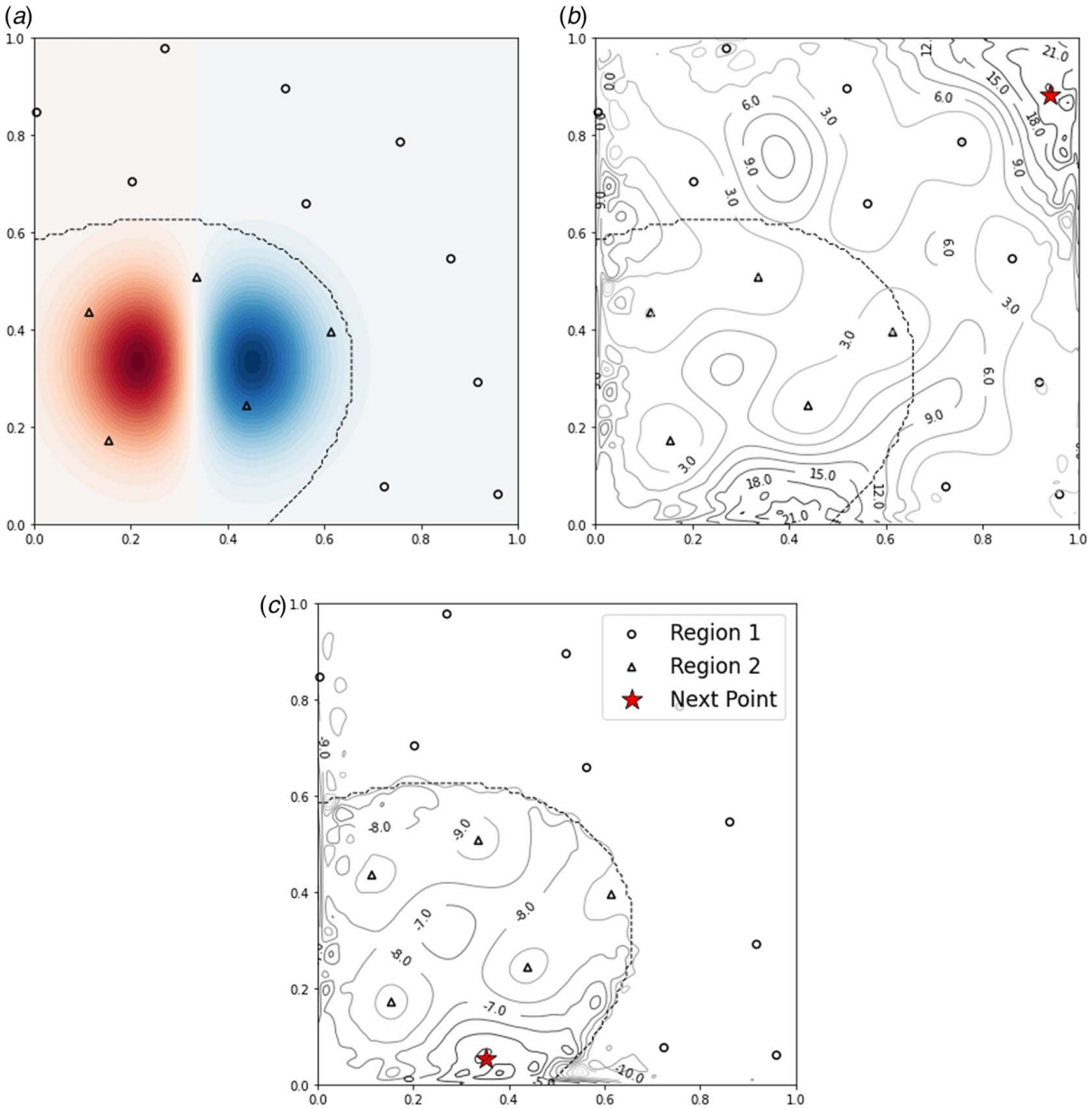

Fig. 4 IMSE criterion contour plots in 2-D simulation: (*a*) Ground truth and initial design points with partitioned regions, (*b*) IMSE of ALC, and (c) PIMSE of PALC

Table 2 Two-dimensional simulation study results

Methods	RMSE	Time (s)
Rand	0.044 (0.017)	—
LHD	0.046 (0.021)	—
ALM	0.060 (0.023)	0.344 (0.081)
PALM	0.019 (0.025)	0.151 (0.026)
ALC	0.039 (0.013)	8.368 (0.161)
PALC-NoG	0.026 (0.004)	5.570 (0.593)
PALC	**0.016 (0.005)**	**1.924 (0.016)**

Note: PALC surpasses the other methods in both predictive accuracy and computational time among the IMSE-based methods, which is denoted in bold.

their fatal affects on the final product. The development of highly accurate predictive model for the shape control is very challenging since the problem is endowed with both heterogeneity and the demanding cost of real experiments. Especially, the stress of composite fuselage is more difficult to predict than deformation [9], we apply our method and other benchmarks to predictive modeling of the stress.

In order to implement our case study cost efficiently, we utilized the FEM model with ANSYS [33], which is well calibrated based on the real experiment. The simulation mimics the real-shape adjustment process that has ten actuators under the fuselage section as shown in Fig. 5(*a*) [34], and the maximum magnitude of actuator's force is 450 lbf. The maximum residual stress on the fuselage section is our interest, which is shown in Fig. 5(*b*) and measured in the psi scale. The ten-dimensional design space is partitioned into three regions with SVC based on finite differences, since a higher number of partitions yields singleton clusters. The Matérn kernel is used for GPs. As the initial dataset, 50 LHD samples are drawn, and additional 30 samples are sequentially obtained from 1,000 LHD points with different active learning strategies. The model evaluation is conducted with a separated testing dataset composed of 100 LHD samples, and mean absolute error (MAE) is used as a metric.

Table 3 summarizes the results of each learning strategy in the case study. PALC surpasses the other methods in both predictive error and the computational time among the IMSE criterion-based methods. Interestingly, except for PALC-NoG, the other active learning methods are worse than two passive learning strategies. The reasons can be summarized as follows. First, the superiority of PALC over ALC tells that partitioning is beneficial in this problem. Second, even though PALM also partitions the design space, the variance criterion thereof is subject to oversampling at boundaries [1]. That is, the variance criterion tends to sample from decision boundaries, which is also preferred in the adjacent subregions due to independence. Moreover, it can be exacerbated when the design space dimension and the number of subregions increase, thereby the decision boundary getting more complexity. In computational times, we can observe that PALC-NoG took more time than ALC.

5.2 Tribocorrosion in Aluminum Alloys. As our second case study, the material loss rate during stress corrosion (i.e., tribocorrosion) in aluminum alloys with six control variables is considered. To test the tribocorrosion resistance of metals, experimental tests and FEM simulations were carried out by scratching the surface of the samples in the corrosive environment [35,36] as shown in Fig. 6. During the tribocorrosion process, the mechanical deformation and the electrochemical processes including active corrosion and passivation work synergistically to cause material degradation. The FEM model calculates the contact mechanics between the indenter and the sample, simulates the wear process as well as the wear-accelerated material dissolution of the corrosion process, and generates the volume loss results. The six control variables for the FEM model are material property descriptors: young's modulus, yield strength, anodic Tafel slope, anodic exchange current density, cathodic Tafel slope, and cathodic exchange current density. The former two govern the mechanical properties, while the latter four determine the corrosion behavior of the alloy. The output of the FEM model is the tribocorrosion rate of the alloy, expressed as volume loss per time.

The surrogate model of the FEM model is constructed to assist the optimal design of alloys with uncertainty quantification and alleviating the high-computational cost of the FEM model. To establish the relationship between material property and tribocorrosion rate, a total of 106 FEM simulations were performed by systematically varying the six control variables. Since scales of variables in the dataset are inconsistent, each variable is normalized to be within a unit interval. For evaluation, relative mean absolute error (RMAE) is used as a metric due to the infinitesimal scale of the output. The RMAE is calculated as follows:

$$\text{RMAE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - h(\mathbf{x}_i)|}{|y_i|}$$

The PGP in this case is composed of three local GPs with the RBF kernel, and the SVC model is used for partitioning based on the finite differences, which provided quite even-sized clusters. Considering the relatively small size of samples, fivefold cross-validation is used. That is, about 84 samples are passed to each

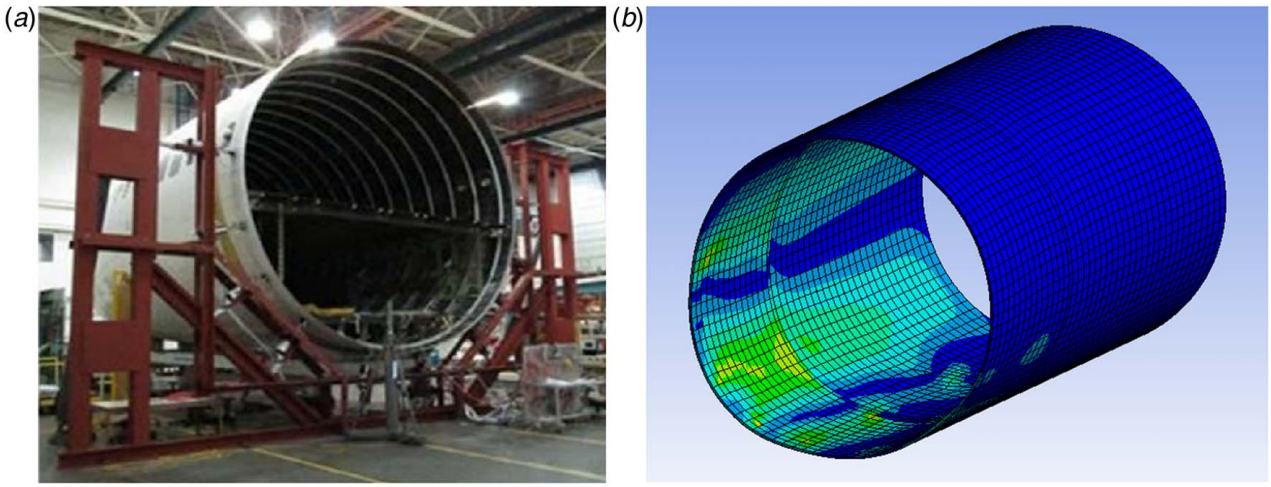

Fig. 5 Shape adjustment of composite fuselage: (*a*) composite fuselage installed upon the fixture with actuators [34] and (*b*) simulated residual stress of composite fuselage in ANSYS

Table 3 Residual stress case study results

Methods	MAE	Time (s)
Rand	3.858	—
LHD	3.319	—
ALM	4.545	2.6×10^{-4}
PALM	11.555	2.5×10^{-4}
ALC	4.693	11.596
PALC-NoG	3.691	12.750
PALC	**3.207**	**9.729**

Note: PALC surpasses the other methods in both predictive error and the computational time among the IMSE criterion-based methods, which is denoted in bold.

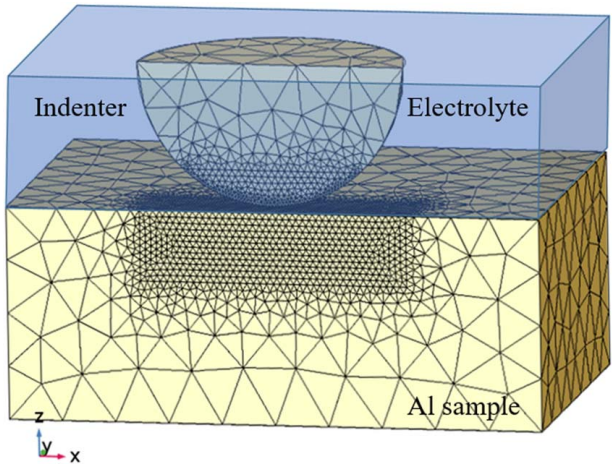

Fig. 6 Schematic tribocorrosion simulation setup

Table 4 Tribocorrosion case study results

Methods	RMAE	Time (s)
Rand	0.028 (0.017)	—
LHD	0.028 (0.017)	—
ALM	0.026 (0.017)	0.014 (0.003)
PALM	0.023 (0.012)	0.104 (0.000)
ALC	0.022 (0.013)	1.283 (0.127)
PALC-NoG	0.022 (0.012)	0.764 (0.018)
PALC	**0.020 (0.013)**	**0.757 (0.017)**

Note: PALC achieves the minimum averaged predictive error and the computational time among the IMSE-based methods, which is denote in bold

model as the candidate set, and the rest of samples are used for the model evaluation. Compared methods are trained up to 50 samples from 20 common initial samples.

Table 4 shows the result of each learning strategy, where the parenthesized numbers are standard deviations from cross-validation. We can observe that PALC achieves the minimum averaged predictive error and the computational time among the IMSE-based methods. Passive learning methods are worse than others, and the variance methods are also worse than the IMSE-based methods. In this case, overall computational times are much lower than the previous case due to the small number of samples in the candidate pool, while the numerical remedies in PALC have significantly reduced the time of ALC.

5.3 Inverse Dynamics of Robot Arm. We apply partitioned active learning to the inverse dynamics problem for a seven degrees-of-freedom robot arm, of which original data are introduced in Ref. [6]. This problem has 21-dimensional input: positions, velocities, and accelerations of seven joints, and seven-joint torques as the output. The dataset contains 44,484 training samples and 4,449 testing samples. We regard the training dataset as the unlabeled data pool, so that we take 30 initial samples with D-optimal design [1] and obtain 30 additional samples from the pool. The testing dataset is referred to as the reference dataset in IMSE-based methods. For model evaluation, standardized MSE (SMSE) is used, which is

$$\mathrm{SMSE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i - f(\mathbf{x}_i)\right)^2}{\mathrm{Var}(Y_{\mathrm{test}})}$$

considering the scale issue. Since we had no specific prior knowledge in the problem, we assigned two partitions according to the finite difference clustering result.

Table 5 Inverse dynamics of robot arm case study results

Methods	SMSE	Time (s)
Rand	1.578	0.474
LHD	1.441	0.551
ALM	1.446	0.973
ALC	1.452	425.740
PALC	**1.093**	**237.672**

Note: PALC can significantly reduce the time of naive IMSE with our proposed method, which is denoted in bold.

Results in Table 5 show that PALC outperforms other methods in the prediction accuracy. The computational time result also shows that PALC can significantly reduce the time of naive IMSE with our proposed method. To compare partitioning methods in high-dimensional setting, we also implemented agglomerative clustering. For the initial dataset, agglomerative clustering was inapplicable when the number of initial dataset was less than the input dimension. For 30 samples, it took 3–4 s, while the proposed partitioning method took only less than 0.1 s. Moreover, for 60 samples, our method took only 0.3 s, while the agglomerative clustering took more than 1200 s with memory overflow error in Voronoi tessellation.

6 Conclusion

Active learning is machine learning that seeks to improve sampling efficiency and lower data collection cost. Existing active learning strategies mainly focus on investigating homogeneous response surfaces, and hence, they are insufficient for reliable and cost-efficient surrogate modeling of heterogeneous systems. This article dedicated establishing an efficient partitioned active leaning strategy that adopts two-step searching schemes based on the PIMSE criterion structure. By partitioning the design space into multiple subregions according to heterogeneity in the target system, the global searching scheme refers to the integrated predictive uncertainties of local GPs to determine the most uncertain subregion. The global searching scheme allows us to reduce the region of interest, thereby not only accelerating the searching speed but also improving the overall learning efficiency. The local searching scheme exploited the chosen local GPs in the global searching phase, so the localized IMSE criterion may provide more relevant information minimizing the interruption of heterogeneous characteristics in other regions. For the numerical perspective of active learning, the following applicable remedies are provided: reducing the number of candidates with the global searching scheme, and the Cholesky factor update, which can be embedded into PALC.

In the simulation and the case study, PALC outperformed the benchmark methods including passive learning, the variance criterion, and the IMSE-based methods. Furthermore, the global searching scheme dramatically improved the performance of PALC by comparing our method to the PIMSE without the global searching

step. Although the proposed method needs some parameter exploration for mean-shift, the partitioning method is computationally much faster in the high-dimensional problems. It would be beneficial to incorporate other types of heterogeneity apart from the provided features in this article, e.g., different distribution family, linear and nonlinear. Diverse case study results imply that our method is also applicable to other domains where heterogeneity exists.

Acknowledgment

C. Lee and X. Yue were supported by the National Science Foundation (Award CMMI-2035038) and the Grainger Frontiers of Engineering Grant Award from the U.S. National Academy of Engineering. W. Cai was supported by the National Science Foundation (Award CMMI-1855651). The authors acknowledge the financial support.

Conflicts of Interest

There are no conflicts of interest.

Data Availability Statement

The data and information that support the findings of this article are freely available.[2,3]

References

[1] Santner, T. J., Williams, B. J., and Notz, W. I., 2018, *The Design and Analysis of Computer Experiments*, Vol. 2, Springer, New York.
[2] Alaeddini, A., Craft, E., Meka, R., and Martinez, S., 2019, "Sequential Laplacian Regularized V-Optimal Design of Experiments for Response Surface Modeling of Expensive Tests: An Application in Wind Tunnel Testing," IISE Trans., **51**(5), pp. 559–576.
[3] Cao, X., Yao, J., Xu, Z., and Meng, D., 2020, "Hyperspectral Image Classification With Convolutional Neural Network and Active Learning," IEEE. Trans. Geosci. Remote. Sens., **58**(7), pp. 4604–4616.
[4] Deisenroth, M. P., Fox, D., and Rasmussen, C. E., 2013, "Gaussian Processes for Data-Efficient Learning in Robotics and Control," IEEE. Trans. Pattern. Anal. Mach. Intell., **37**(2), pp. 408–423.
[5] Yue, X., Wen, Y., Hunt, J. H., and Shi, J., 2020, "Active Learning for Gaussian Process Considering Uncertainties With Application to Shape Control of Composite Fuselage," IEEE Trans. Autom. Sci. Eng., **18**(1), pp. 36–46.
[6] Rasmussen, C., and Williams, C., 2006, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA.
[7] Ghoreishi, S. F., and Imani, M., 2021, "Bayesian Surrogate Learning for Uncertainty Analysis of Coupled Multidisciplinary Systems," ASME J. Comput. Inf. Sci. Eng., **21**(4), p. 041009.
[8] Hyer, M. W., and White, S. R., 2009, *Stress Analysis of Fiber-Reinforced Composite Materials*, DEStech Publications, Inc, Lancaster, PA.
[9] Lee, C., Wu, J., Wang, W., and Yue, X., 2020, "Neural Network Gaussian Process Considering Input Uncertainty for Composite Structures Assembly," IEEE/ASME Transact. Mechatron, pp. 1–1.
[10] Kim, H.-M., Mallick, B. K., and Holmes, C., 2005, "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes," J. Am. Stat. Assoc., **100**(470), pp. 653–668.
[11] Heaton, M. J., Christensen, W. F., and Terres, M. A., 2017, "Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering From Finite Differences," Technometrics, **59**(1), pp. 93–101.
[12] Gramacy, R. B., and Lee, H. K. H., 2008, "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," J. Am. Stat. Assoc., **103**(483), pp. 1119–1130.
[13] Konomi, B., Karagiannis, G., Sarkar, A., Sun, X., and Lin, G., 2014, "Bayesian Treed Multivariate Gaussian Process With Adaptive Design: Application to a Carbon Capture Unit," Technometrics, **56**(2), pp. 145–158.
[14] Gramacy, R. B., and Lee, H. K., 2009, "Adaptive Design and Analysis of Supercomputer Experiments," Technometrics, **51**(2), pp. 130–145.
[15] Pope, C. A., Gosling, J. P., Barber, S., Johnson, J. S., Yamaguchi, T., Feingold, G., and Blackwell, P. G., 2021, "Gaussian Process Modeling of Heterogeneity and Discontinuities Using Voronoi Tessellations," Technometrics, **63**(1), pp. 53–63.
[16] Seo, S., Wallat, M., Graepel, T., and Obermayer, K., 2000, "Gaussian Process Regression: Active Data Selection and Test Point Rejection," Mustererkennung 2000: 22. DAGM-Symposium, Kiel, Germany, Sept. 13–15, Springer, pp. 27–34.
[17] Chen, X., and Zhou, Q., 2017, "Sequential Design Strategies for Mean Response Surface Metamodeling Via Stochastic Kriging With Adaptive Exploration and Exploitation," Eur. J. Oper. Res., **262**(2), pp. 575–585.
[18] Lee, C., Wang, X., Wu, J., and Yue, X., 2022, "Failure-Averse Active Learning for Physics-Constrained Systems," IEEE Trans. Autom. Sci. Eng., pp. 1–12.
[19] Ghassemi, P., and Chowdhury, S., 2020, "An Extended Bayesian Optimization Approach to Decentralized Swarm Robotic Search," ASME J. Comput. Inf. Sci. Eng., **20**(5), p. 051003.
[20] AlBahar, A., Kim, I., and Yue, X., 2022, "A Robust Asymmetric Kernel Function for Bayesian Optimization, With Application to Image Defect Detection in Manufacturing Systems," IEEE Trans. Autom. Sci. Eng., **19**(4), pp. 3222–3233.
[21] Pasolli, E., and Melgani, F., 2011, "Gaussian Process Regression Within an Active Learning Scheme," 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, Canada, July 24–29, IEEE, pp. 3574–3577.
[22] Käding, C., Rodner, E., Freytag, A., Mothes, O., Barz, B., Denzler, J., and AG, C. Z., 2018, "Active Learning for Regression Tasks With Expected Model Output Changes," BMVC 2018, Newcastle, UK, Sept. 3–6, BMVC, p. 103.
[23] Erickson, C. B., Ankenman, B. E., Plumlee, M., and Sanchez, S. M., 2018, "Gradient Based Criteria for Sequential Design," 2018 Winter Simulation Conference (WSC), Gothenburg, Sweden, Dec. 9–12, IEEE, pp. 467–478.
[24] Marmin, S., Ginsbourger, D., Baccou, J., and Liandrat, J., 2018, "Warped Gaussian Processes and Derivative-Based Sequential Designs for Functions With Heterogeneous Variations," SIAM/ASA J. Uncertain. Quantification, **6**(3), pp. 991–1018.
[25] Kim, B., Lee, Y., and Choi, D.-H., 2009, "Construction of the Radial Basis Function Based on a Sequential Sampling Approach Using Cross-Validation," J. Mech. Sci. Technol., **23**(12), pp. 3357–3365.
[26] Liu, H., Ong, Y.-S., and Cai, J., 2018, "A Survey of Adaptive Sampling for Global Metamodeling in Support of Simulation-Based Complex Engineering Design," Struct. Multidiscipl. Optim., **57**(1), pp. 393–416.
[27] Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H., 2016, "Non-Stationary Gaussian Process Regression With Hamiltonian Monte Carlo," Artificial Intelligence and Statistics, PMLR, Cadiz, Spain, May 7–11, PMLR, pp. 732–740.
[28] Rasmussen, C. E., and Ghahramani, Z., 2002, "Infinite Mixtures of Gaussian Process Experts," NIPS 2001, Vancouver, Canada, Dec. 3–8.
[29] Park, C., and Apley, D., 2018, "Patchwork Kriging for Large-Scale Gaussian Process Regression," J. Mach. Learn. Res., **19**(1), pp. 269–311.
[30] Aurenhammer, F., and Klein, R., 2000, "Voronoi Diagrams.," Handb. Comput. Geometry, **5**(10), pp. 201–290.
[31] Comaniciu, D., and Meer, P., 2002, "Mean Shift: A Robust Approach Toward Feature Space Analysis," IEEE. Trans. Pattern. Anal. Mach. Intell., **24**(5), pp. 603–619.
[32] Comaniciu, D., 2003, "An Algorithm for Data-Driven Bandwidth Selection," IEEE. Trans. Pattern. Anal. Mach. Intell., **25**(2), pp. 281–288.
[33] Wang, Y., Yue, X., Tuo, R., Hunt, J. H., and Shi, J., et al.2020, "Effective Model Calibration via Sensible Variable Identification and Adjustment With Application to Composite Fuselage Simulation," Ann. Appl. Stat., **14**(4), pp. 1759–1776.
[34] Wen, Y., Yue, X., Hunt, J. H., and Shi, J., 2018, "Feasibility Analysis of Composite Fuselage Shape Control via Finite Element Analysis," J. Manuf. Syst., **46**, pp. 272–281.
[35] Wang, K., Wang, Y., Yue, X., and Cai, W., 2021, "Multiphysics Modeling and Uncertainty Quantification of Tribocorrosion in Aluminum Alloys," Corros. Sci., **178**, p. 109095.
[36] Wang, K., and Cai, W., 2021, "Modeling the Effects of Individual Layer Thickness and Orientation on the Tribocorrosion Behavior of AL/CU Nanostructured Metallic Multilayers," Wear, **477**, p. 203849.

[2]https://doi.org/10.24433/CO.5741905.v2
[3]https://github.com/cheolheil/ALIEN