ASME Journal of Mechanical Design
Online journal at:
https://asmedigitalcollection.asme.org/mechanicaldesign

Data-Driven Global Sensitivity Analysis of Variable Groups for Understanding Complex Physical Interactions in Engineering Design

Tuba Dolar
Department of Mechanical Engineering,
Northwestern University,
2145 Sheridan Road,
Evanston, IL 60208
e-mail: tubadolar@u.northwestern.edu

Doksoo Lee
Department of Mechanical Engineering,
Northwestern University,
2145 Sheridan Road,
Evanston, IL 60208
e-mail: dslee@u.northwestern.edu

Wei Chen[1]
Department of Mechanical Engineering,
Northwestern University,
2145 Sheridan Road,
Evanston, IL 60208
e-mail: weichen@northwestern.edu

In engineering design, global sensitivity analysis (GSA) is used for analyzing the effects of inputs on the system response and is commonly studied with analytical or surrogate models. However, such models fail to capture nonlinear behaviors in complex systems and involve several modeling assumptions. Besides model-focused methods, a data-driven GSA approach, rooted in interpretable machine learning, would also identify the relationships between system components. Moreover, a special need in engineering design extends beyond performing GSA for input variables individually, but instead evaluating the contributions of variable groups on the system response. In this article, we introduce a flexible, interpretable artificial neural network model to uncover individual as well as grouped global sensitivity indices for understanding complex physical interactions in engineering design problems. The proposed model allows the investigation of the main effects and second-order effects in GSA according to functional analysis of variance (FANOVA) decomposition. To draw a higher-level understanding, we further use the subset decomposition method to analyze the significance of the groups of input variables. Using the design of a programmable material system (PMS) as an example, we demonstrate the use of our approach for examining the impact of material, architecture, and stimulus variables as well as their interactions. This information lays the foundation for managing design space complexity, summarizing the relationships between system components, and deriving design guidelines for PMS development. [DOI: 10.1115/1.4064633]

Keywords: data-driven design, global sensitivity analysis, interpretable machine learning, artificial neural networks, grouped global sensitivity indices, sensitivity analysis for design

1 Introduction

Global sensitivity analysis (GSA) is widely used in engineering design to study the inner workings of complex systems. GSA identifies the contribution of system inputs to the uncertainty of the output by analyzing the impacts of design inputs on the model response. Model verification, model simplification, and establishment of research priorities for identifying the critical model inputs are examples that benefit from GSA.

Numerous statistical methods have been developed for studying the relationship between the model inputs and outputs. Sensitivity analysis is a task related to uncertainty quantification and has been extensively studied in statistics. The first historical approaches to sensitivity analysis focused on revealing the impact of small input perturbations on system response, which is referred to as local sensitivity analysis (LSA) [1]. Following that, methods that consider the variation of the entire model parameters were developed in statistical frameworks under GSA approaches [1]. Among them, regression coefficients are used for sensitivity analysis purposes where a simple linear model is first fit, then its regression coefficients are regarded as sensitivity indices [2]. Variance-based methods decompose the output variance into terms caused by combinations of input variable and input variable groups. Here, sensitivity is assessed with the amount of output variance explained by an input [3]. Some other approaches include design of experiments, graphical methods, Fourier amplitude sensitivity test (FAST), mutual information index, mathematical approximation strategies such as polynomial chaos expansion (PCE), and so on [4].

Despite the availability of diverse techniques, statistical methods use analytical or surrogate models to study the relationships between system inputs and outputs. However, model-focused approaches involve critical challenges created by the modeling assumptions and limitations. When working with a linear regression model, for example, we assume that the true model form is linear,

[1]Corresponding author.

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received October 28, 2023; final manuscript received January 23, 2024; published online March 5, 2024. Assoc. Editor: Xueguan Song.

residuals are normally distributed, they have equal variance, and samples are independent. Such conditions must be satisfied for the model to be valid, restricting the flexibility to accurately identify complex behaviors. Most real-world problems consist of numerous subsystems interacting with each other, leading to an increase in model size and complexity. In such cases, system models are difficult to define and even if they are built, they fail to capture nonlinear behaviors. On the contrary, researchers can conduct physical experiments or run simulations to collect data about a system. With the availability of sufficient data, machine learning models usually allow for obtaining accurate models of complex systems where the physics behind the model behavior is not analytically unknown. Along with being flexible, machine learning models can also be employed for GSA purposes.

Even though black-box machine learning models do not serve the objectives of GSA, interpretable machine learning focuses on transforming black-box models into glass-box models which can provide insights from a sensitivity point of view. Interpretable models offer justifications behind model predictions as "Interpretability is the degree to which a human can understand the cause of a decision" [5]. The more interpretable a model is, the easier for researchers to understand its inner workings, which perfectly aligns with the purposes of GSA. Similar to sensitivity analysis goals, various interpretable machine learning techniques are available for local and global analysis. Local interpretability explains why an individual prediction is made whereas global interpretability describes the entire model behavior. In the engineering design context, interpretable machine learning would allow researchers to identify the relationships between model variables, eliminating the necessity of model-based sensitivity analysis implementations. In this article, we introduce a data-driven approach rooted in interpretable machine learning for performing GSA in engineering design.

Apart from a data-driven GSA approach for evaluating individual variables, a special need in engineering design problems is extending GSA results for assessing the contributions of variable groups on system response. Dividing the independent variables into groups and studying the contributions of these is valuable, for example in cases where the model is complex with many variables or where there are explicitly meaningful variable groups. One example would be robust design where researchers aim at designing systems that meet the performance requirement regardless of the many sources of variation. In robust design, two broad problems involve: first, minimization of the performance variations caused by variations in noise factors and second, minimization of the performance variations caused by control factors [6]. In this framework, it is valuable to evaluate the sensitivity indices for groups of uncontrollable parameters and design variables. Studying the importance and effect of these two groups helps with mitigating the effects of the uncontrollable parameters. A second example of motivation is the design of programmable material systems (PMS) manufactured with smart materials that are responsive to an external stimulus such as magnetic field, temperature, or humidity [7]. This property brings the opportunity of programming PMS to change their shapes and dimensions for performing sophisticated functions, controlled drug release for example. For such complex structures with high-dimensional representations of spatially diverse material composition, topological architecture, and external stimulus, data-driven GSA of variable groups allows the investigation of the role of material, architecture, and stimulus. This information then can be used as input for prioritizing the design efforts, considering that material design, architecture design, and stimulus design belong to different domains within multidisciplinary design efforts. In cases where for instance material variables are identified as the biggest contributors, research efforts and resources could be centered around the material design domain. Furthermore, if the interactions between material and architecture variable groups stand out, collaborative design endeavors between these two teams will be critical for achieving the desired performance.

Combining these two objectives, performing data-driven GSA, and further extending the analysis to variable groups, we propose employing an interpretable neural network model. The interpretable neural network model is already introduced in the literature [8] and allows the investigation of the importance of the main effects and second-order effects of the input variables. To draw a higher-level understanding, we further propose a subset decomposition method [9] to be combined with the interpretable model for analyzing the significance of groups of input variables. The synthesis of these two techniques allows us to:

(1) improve an interpretable neural network analysis, which works with individual input variables, to further interpret variable groups,
(2) reveal the complex physical interactions in engineering design problems,
(3) represent an application and the benefits of data-driven global sensitivity analysis in an engineering design context.

The remainder of the article is organized as follows. In Sec. 2, we start by reviewing the relevant studies and provide background information. Following that, Sec. 3 introduces the interpretable neural network model which is then combined with the subset decomposition approach as a contribution of this work to serve the need for the GSA of variable groups as shown in Fig. 1. Then, Sec. 4 demonstrates our approach in a mathematical example problem and a programmable photonic metasurface design problem where we elucidate the complex relationships and compile design rules. Finally, Sec. 5 concludes the article and discusses future work.

2 Background

2.1 Interpretable Machine Learning. Interpretable machine learning resolves the disadvantages of black-box models and allows an understanding of how a machine-learning model works. Research on interpretability involves two types: post hoc and intrinsic which is determined by whether interpretability is achieved by using additional methods after a predictive model is obtained or whether it is achieved during the training phase [10].

Intrinsically interpretable models mostly have simple structures which make it easy to evaluate model behavior. For example, sparse linear models and decision trees are easy to interpret without any additional effort. The coefficients of linear models can be referred to as feature importance. As for the decision tree models, measuring how much a split reduces the Gini index compared to the parent node reveals the importance of that variable.

Post hoc interpretability includes adopting an external method after the model training is completed. This follow-up interpretability analysis after obtaining a model can aim to understand a local decision or draw global inferences. Post hoc interpretability tools are model-agnostic and can be adapted to any model. Some examples of post hoc methods for global interpretability include partial dependence plots (PDP) [11], accumulated local effects (ALE) [12] plots, and permutation feature importance (PFI) [13]. PDP and ALE are visual tools that show how one or two features affect model prediction. PFI evaluates an input's importance by varying it and observing the change in the model's prediction error; the more important the input is, the higher the model error becomes. Besides, post hoc interpretability of local decisions can be achieved with individual conditional expectation (ICE) [14] plots, local interpretable model-agnostic explanations (LIME) [15], and Shapley additive explanations (SHAP) [16]. ICE plots are equivalent to PDP with the only difference being constructed for individual data points. An ICE plot consists of separate lines per instance showing how the instance's prediction changes with respect to a feature. LIME involves training a simple local surrogate model around the area of interest to explain the individual prediction. SHAP is adopted from game theory and explains the contribution of each input on the model prediction for individual data points.

A tradeoff between predictive performance and interpretability is reported in the literature [17–20]. That is, intrinsically interpretable

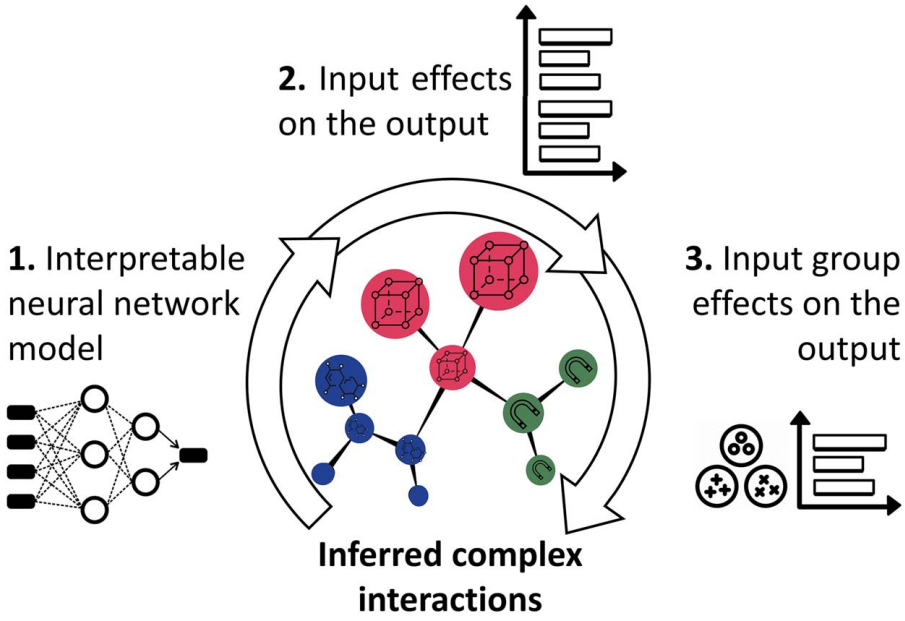


Fig. 1 Steps of the proposed method for identifying contributions of the variable groups on the system response

models limit the model complexity which hurts the predictive accuracy. In such cases where predictive performance is more critical than interpretation, post hoc methods are favored. Still, intrinsically interpretable models are advantageous as they extract sensitivity indices of the input variables together with providing a predictive model. Considering these benefits, an intrinsically interpretable modeling approach is adopted in this study.

2.2 Sensitivity Analysis With Artificial Neural Networks. The proposed approach in this article employs artificial neural networks (ANNs) for identifying the importance of model inputs. ANNs provide distinguished predictive advantages for detecting complex nonlinear relationships between model inputs and outputs. Nevertheless, they have complex model structures with varying numbers of layers, all sorts of connections, and numerous neurons with many weight and bias parameters. Consequently, finding methods for revealing how an ANN model uses the inputs for predicting the output is an expanding research area.

Several existing methods address ANN weights as being partially analogous to coefficients of a linear model and use these for revealing the input contributions. Simply dividing the sum square of the weights of the input variable of interest by the sum square of the weights of all input variables is an intuitive approach in this regard [21]. Similarly, weights from the input variable through the hidden layers up until the output variable can be tallied and scaled relative to all input variables [22]. Adding noise to each input variable one at a time is also useful for assessing the change in a chosen error metric which signals the input importance [23]. Moreover, partial derivatives of the ANN outputs with respect to input neurons is also informative from a GSA point of view [24]. Some other approaches construct an ANN model sequentially where inputs and their weights are added or eliminated. Tracing the change in a specified error metric indicates the importance of the input variable [25]. As introduced previously, PDP, ICE, and LIME visualization tools can also be utilized with ANN as the surrogate model for revealing the input variable contributions.

2.3 Global Sensitivity Analysis of Variable Groups. For cases when identifying the contribution of groups of variables on the model prediction is meaningful, several statistical methods are offered in the literature. Sobol indices [26], a very well-known variance-based sensitivity analysis method, can handle either individual inputs or sets of inputs as the output variance can be decomposed with respect to input groups. Morris's method of grouping [27] also allows for analyzing variable groups by varying the variables within a group simultaneously along a trajectory and then observing the change in the system response. However, this method fails to distinguish low and high-order interactions. Derivative-based global sensitivity measures (DSGM) [28] calculate the average of local derivatives of the variables from the same group. This approach involves working with gradients, which is not applicable to problems with categorical variables. The approach introduced in this article further extends GSA results to variable groups while eliminating many of the challenges raised by these methods.

3 Methods

Aiming to acquire the benefits of data-driven design, we propose combining two methods for conducting data-driven GSA for variable groups. Our approach first starts with obtaining an intrinsically interpretable deep neural network model of a given dataset instead of first training a model and then using post hoc interpretability tools. This intrinsically interpretable system model delivers the sensitivity indices of individual input variables. Then subset decomposition is introduced to obtain the sensitivity indices of variable groups. In the end, this workflow enables revealing the complex interactions between system components and accordingly manages the design space complexity as described in Fig. 2.

The proposed approach allows the following:

(1) utilizing data-driven design when an analytical or surrogate model is unavailable or expensive,
(2) managing large datasets and high-dimensional problems,
(3) avoiding any model limitations as the approach is flexible to capture nonlinear system behavior,
(4) working with mixed input spaces consisting of both continuous and categorical variables,
(5) working with both regression and classification tasks,
(6) performing both GSA and LSA,
(7) visually interpreting the input variable contributions.

3.1 Functional Analysis of Variance Decomposition. The suggested approach is based on functional analysis of variance

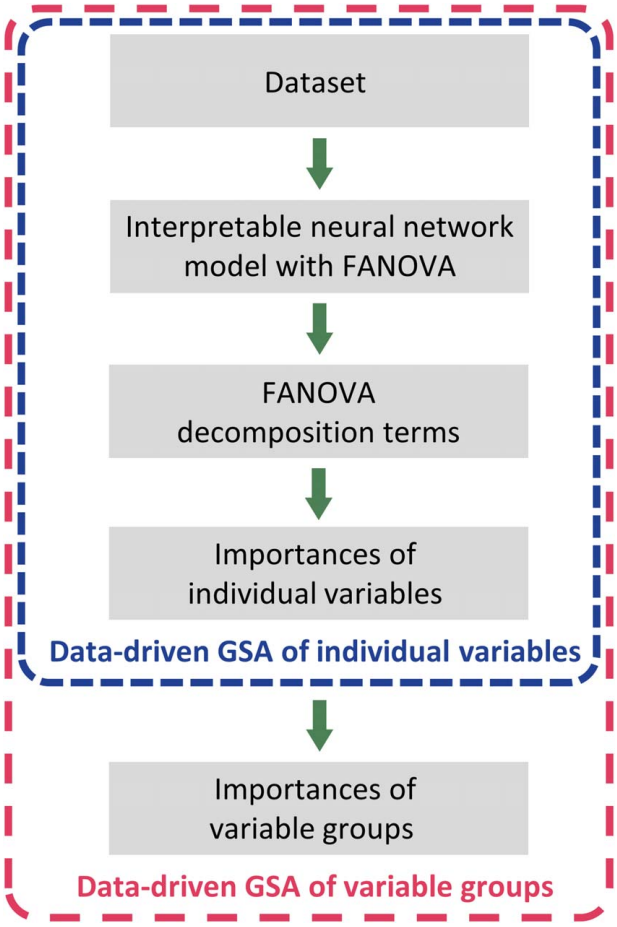


Fig. 2 Flowchart of the proposed data-driven global sensitivity analysis approach for variable groups

decomposition (FANOVA) where the variance of a function is decomposed into terms attributable to input variables and their relationships [29]. In this section, first, a mathematical example is used to introduce the approach, and then FANOVA representation terms are presented.

Considering a simple function defined in Eq. (1) as an example, the decomposed form of $f(\boldsymbol{x})$ contains the 4 terms in Eq. (2). Among these, f_0 represents a constant value, f_1 refers to what happens in $f(\boldsymbol{x})$ when x_1 is varied, similarly f_2 shows what happens in $f(\boldsymbol{x})$ when x_2 is varied, and f_{12} is the case when x_1 and x_2 vary together.

$$f(\boldsymbol{x}) = x_1 x_2 \qquad x_i \in \{0, 1\},\ \forall i \tag{1}$$

$$f(\boldsymbol{x}) = f_0 + f_1 + f_2 + f_{12} \tag{2}$$

The decomposition terms involve the calculation of conditional expected values which are provided for the example problem in Eqs. (3)–(6).

$$f_0 = E[f(\boldsymbol{x})] = \frac{1}{4} \tag{3}$$

$$f_1 = E[f(\boldsymbol{x})|x_1] - f_0 = \frac{x_1}{2} - \frac{1}{4} \tag{4}$$

$$f_2 = E[f(\boldsymbol{x})|x_2] - f_0 = \frac{x_2}{2} - \frac{1}{4} \tag{5}$$

$$f_{12} = E[f(\boldsymbol{x})|x_1, x_2] - f_1 - f_2 - f_0 = x_1 x_2 - \frac{x_1}{2} - \frac{x_2}{2} + \frac{1}{4} \tag{6}$$

After obtaining the decomposed representation of $f(\boldsymbol{x})$, n being the number of samples, function variance is calculated using the decomposition terms and the orthogonality between them in Eq. (7).

$$\mathrm{Var}(f(\boldsymbol{x})) = \frac{(f(\boldsymbol{x}) - E[f(\boldsymbol{x})])^2}{n-1} = \frac{\{(f_0 + f_1 + f_2 + f_{12}) - f_0\}^2}{n-1} = \frac{f_1^2 + f_2^2 + f_{12}^2}{n-1} \tag{7}$$

Then, the sensitivity indices S_1, S_2, and S_{12} represent how much of the total variance $\mathrm{Var}(f(\boldsymbol{x}))$ is caused by the individual terms f_1^2, f_2^2, and f_{12}^2 as shown in Eqs. (8)–(10).

$$S_1 = \frac{f_1^2}{f_1^2 + f_2^2 + f_{12}^2} \tag{8}$$

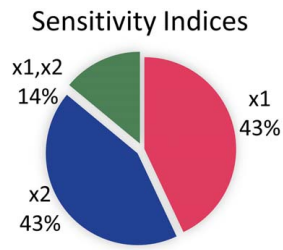
$$S_2 = \frac{f_2^2}{f_1^2 + f_2^2 + f_{12}^2} \tag{9}$$

$$S_{12} = \frac{f_{12}^2}{f_1^2 + f_2^2 + f_{12}^2} \tag{10}$$

Results of GSA with FANOVA for the example problem in Eq. (1) provide the pie chart in Fig. 3. The contributions of x_1 and x_2, which are named as the main effects in the sensitivity analysis studies, are equal and account for 43% of the variance in the output whereas the second-order effect of x_1 and x_2 have 14% importance. Analyzing the main effects is important for not just eliminating insignificant model variables but also summarizing the impact of model inputs, x_1 and x_2 having the same impact in Eq. (1) for example. While main effects refer to the controlled variables, second-order effects arise because of the interactions between these variables, for which the researchers do not have direct control over. In cases where the main effect of a variable is not significant, it can be removed from the model. However, if a second-order effect of the same variable holds substantial weight, its exclusion creates a change caused by not only the negligible main effect but also the critical second-order effect. Accordingly, analyzing main and second-order effects together is essential for summarizing model behavior.

Following this example, FANOVA representation of any model can be expressed with Eq. (11) where for $\forall i = \{1,\ 2,\ \ldots,\ M\}$, x_i is an independent random variable with probability density functions $p_i(x_i)$, $\varphi_i(x_i)$ is main effects, and $\varphi_{i_1 i_2}(x_{i_1}, x_{i_2})$ is second-order effects.

$$f(x_1,\ \ldots,\ x_M) = f_0 + \sum_{i=1}^{M} \varphi_i(x_i) + \sum_{i_1=1}^{M} \sum_{i_2=i_1+1}^{M} \varphi_{i_1 i_2}(x_{i_1}, x_{i_2}) + \ldots + \varphi_{1\ldots M}(x_1, \ldots x_M) \tag{11}$$



Main effects: x1 = x2 = 0.43, *Second order effects:* x1,x2 = 0.14, *Total effects:* x1 = x2 = 0.57

Fig. 3 Global sensitivity analysis results of the example problem

The model is decomposed into a constant mean, main effects, second-order effects, and so on [29]. Preferably, a few low-order terms, Eqs. (12)–(14), will be sufficient to approximate f, eliminating the necessity of complex decomposed model representations.

$$f_0 = \int f(\boldsymbol{x}) \prod_{i=1}^{M} [\, p_i(x_i) dx_i] \tag{12}$$

$$\varphi_i(x_i) = \int f(\boldsymbol{x}) \prod_{j \neq i} [p_j(x_j) d\boldsymbol{x}_j] - f_0 \tag{13}$$

$$\varphi_{i_1 i_2}(x_{i_1}, x_{i_2}) = \int f(\boldsymbol{x}) \prod_{j \neq i_1, i_2} [p_j(x_j) d\boldsymbol{x}_j] - \varphi_{i_1}(x_{i_1}) - \varphi_{i_2}(x_{i_2}) - f_0 \tag{14}$$

3.2 Functional Analysis of Variance Decomposition of an Artificial Neural Network Model. The interpretable neural network model is a generalized additive model based on FANOVA where the dependent variable is the sum of a combination of variables as shown in Eq. (15) [8]. The method captures the main effects, $h_j(x_j)$ in Eq. (16), as well as the second-order effects, $f_{jk}(x_j,\ x_k)$ in Eq. (17), where is a μ constant value and $F(x_j)$, $F(x_j,\ x_k)$, and $F(\boldsymbol{x})$ are the respective cumulative distributions.

$$g(E[y|\boldsymbol{x}]) = \mu + \sum_{j \varepsilon S_1} h_j(x_j) + \sum_{(j,k) \varepsilon S_2} f_{jk}(x_j,\ x_k) \tag{15}$$

$$\int h_j(x_j) dF(x_j) = 0,\ \ \forall j \in S_1 \tag{16}$$

$$\int f_{jk}(x_j,\ x_k) dF(x_j,\ x_k) = 0,\ \ \forall (j,k) \varepsilon S_2 \tag{17}$$

$$\int h_j(x_j) f_{jk}(x_j,\ x_k) dF(\boldsymbol{x}) = 0 \tag{18}$$

Equation (16) ensures that the sum of the main effects cancel each other out; so that when some variables have positive impact on the output, others have negative impact to enforce $g(E[y|\boldsymbol{x}]) = \mu$ after assessing all variable impacts. Similarly, Eq. (17) represents the same constraint for the second-order effects where all second-order effects negate one another in the end. Equation (18) enforces the orthogonality of the main and second-order effects which is essential for differentiating whether the impact of a variable is caused by its main effect or interaction effect. As an example, main effect x_1 and second-order effect x_1, x_2 both contain the same variable x_1. In this situation, it is critical to identify which one accounts for the contribution of x_1. This constraint imposes that main and second-order effects are independent by being orthogonal.

As explained in Sec. 3.1, the calculation of main and second-order effects involve conditional expected values which are approximated with ANN in the introduced method. The ANN model starts with μ, the average of the output value then captures the main effects, and in the second stage, the second-order effects as demonstrated in Fig. 4.

Model training starts with capturing the main effects for which fully connected subnetworks are trained simultaneously between each individual input variable and the output. Subnetworks of the input variables are then introduced to the model sequentially, and the change in the loss function is traced. Input variables are ranked according to their impact on the loss function and added to the model in descending order. After identifying the significant main effects in this first stage, a similar approach is adopted for the second-order effects where fully connected subnetworks are fitted between two input variables and the residuals after the first

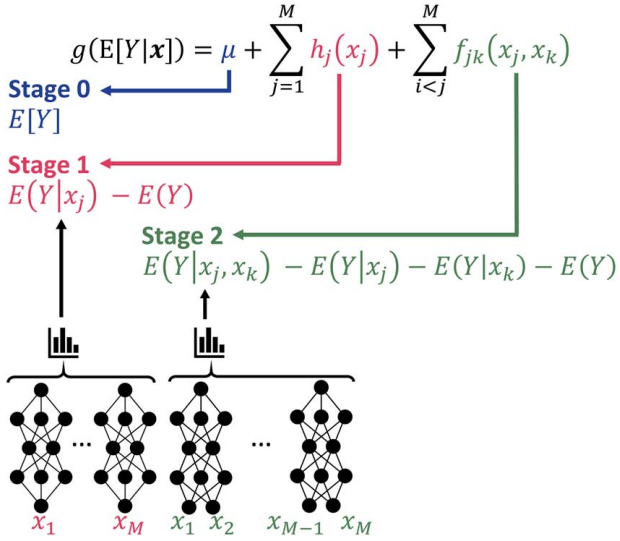


Fig. 4 Schematic of the interpretable neural network model

stage. These subnetworks with pairs of input variables in the input layer are added to the model sequentially and change in the loss function is monitored. Input variable pairs are ranked with respect to the change they bring upon the loss function and added to the model in descending order. After identifying $h_j(x_j)$ and $f_{jk}(x_j,\ x_k)$ terms to be included in the ANN model, all network parameters are fine-tuned.

For enhancing interpretability and limiting model complexity, the method employs several constraints. First, insignificant main and second-order effects are pruned to make sure that the ANN model contains only critical terms. Furthermore, a second-order effect is kept in the model only if at least one of its parent main effects is included in the model. Finally, the main effects and their corresponding child effects are not correlated as dictated in Eq. (18).

The constrained optimization problem in Eq. (19) is solved by obtaining $h_j(x_j)$ and $f_{jk}(x_j,\ x_k)$ terms where $\boldsymbol{\theta}$ is the model parameters, $l(\boldsymbol{\theta})$ is a loss defined by the corresponding regression or classification task, $\Omega(x_j,\ x_k) = \left|\frac{1}{n}\sum h_j(x_j) f_{jk}(x_j,\ x_k)\right|$ n being the number of samples, and λ is the coefficient for the orthogonality constraint.

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \mathcal{L}_\lambda(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) + \lambda \sum_{j \varepsilon S_1} \sum_{(j,k) \varepsilon S_2} \Omega(x_j,\ x_k), \\ &\text{s.t.} \int h_j(x_j) dF(x_j) = 0,\ \ \forall j \in S_1, \\ &\int f_{jk}(x_j,\ x_k) dF(x_j,\ x_k) = 0,\ \ \forall (j,k) \varepsilon S_2 \end{aligned} \tag{19}$$

$h_j(x_j)$ and $f_{jk}(x_j,\ x_k)$ are then used for quantifying the importance of main and second-order effects as shown in Eqs. (20) and (21) where n is the number of samples, $D(h_j) = \frac{1}{n-1}\sum h_j^2(x_j)$, and $D(f_{jk}) = \frac{1}{n-1}\sum f_{jk}^2(x_j,\ x_k)$.

$$IR(j) = \frac{D(h_j)}{\sum_{j \varepsilon S_1} D(h_j) + \sum_{(j,k) \varepsilon S_2} D(f_{jk})} \tag{20}$$

$$IR(j,k) = \frac{D(f_{jk})}{\sum_{j \varepsilon S_1} D(h_j) + \sum_{(j,k) \varepsilon S_2} D(f_{jk})} \tag{21}$$

3.3 Variable Set Decomposition. We employ subset decomposition [9], which is a variance-based sensitivity analysis method similar to the FANOVA decomposition for individual variables, for assessing the impact of variable groups. In this study, we cover problems where meaningful variable groups are already

defined, that is by predefining variable groups based on the physical nature of these variables and application of interest. Assuming that the groups are statistically independent, we use the same logic for variance decomposition of the input variable groups as shown in Eq. (22) where "^" refers to the decomposition items belonging to the same group.

$$f(\boldsymbol{x}) = f_0 + \sum \hat{\varphi}_{U_i}(x_{U_i}) + \sum_{i_1=1}^{r} \sum_{i_2=i_1+1}^{r} \hat{\varphi}_{U_{i_1}U_{i_2}}(x_{U_{i_1}U_{i_2}},\ x_{U_{i_1}U_{i_2}}) + \hat{\varphi}_{U_1\ldots U_T}(x_{U_1},\ \ldots,\ x_{U_T}) \tag{22}$$

Total variation now can be decomposed in Eq. (23) into the sum of variances caused by the groups.

$$V = \sum \hat{V}_{U_i} + \sum_{i_1<i_2} \hat{V}_{U_{i_1}U_{i_2}} + \ldots + \hat{V}_{U_1\ldots U_T} \tag{23}$$

The importance of each subset then represents how much of the total variability each group accounts for as expressed in Eq. (24). Obtained group importance values correspond to grouped global sensitivity indices which can be used for model interpretability with respect to input groups.

$$\hat{S}_{U_{i_1}\ldots U_{i_s}} = \frac{\hat{V}_{U_{i_1}\ldots U_{i_s}}}{V} \tag{24}$$

4 Results

4.1 Mathematical Example. As the first example, we analyze an analytical function with 10 input variables shown in Eq. (25). The defined function is a high-dimensional problem with 10 input variables, involving terms with a variety of interactions of varying shapes and strengths between them. The problem of interest is a complex function with different mathematical operations including trigonometric, logarithmic, and exponential calculations. The goal of applying the proposed data-driven GSA for variable groups in this problem is to analyze the effect of different mathematical operation terms on the function output.

$$f(\boldsymbol{x}) = \tanh(x_1x_2 + x_3x_4)\sqrt{|x_5|} + 0.3e^{x_5+x_6} + \log\left((x_6x_7x_8)^2 + 1\right) + 2x_5x_{10} + \frac{1}{|3x_9| + |3x_{10}|} \tag{25}$$

The importance of the main effects is not easy to predict before the analysis as it highly depends on the sampling distribution of the variables as well as the respective mathematical operation type. Still, second-order effects of variables contained within the same mathematical operation, such as x_5 and x_6 from the exponential term and x_6, x_7, and x_8 from the logarithmic term, are expected before performing the analysis.

4.1.1 Dataset Construction. A dataset of 10,000 samples is constructed with the Latin hypercube sampling method for the function $f(\boldsymbol{x})$. All data points are sampled from the same distribution $U(-1,\ 1)$ over a domain $Z = X \times Y$ where $X \in \mathbb{R}^{10}$, $Y \in \mathbb{R}$.

4.1.2 Data-Driven Global Sensitivity Analysis of Individual Variables. We build an interpretable ANN model for modeling Eq. (25) using the toolbox [30] established by the interpretable ANN model developers [8]. A ratio of 0.8/0.2 is used for randomly splitting the dataset into training and testing sets. In the network architecture, we consider the main effects and the most critical first 20 s second-order effects. Each subnetwork has 5 ReLU hidden layers with 40 nodes per layer. The batch size is 512 while the maximum number of epochs for the main effects, second-order effects, and model tuning is 1000. The learning rates of the Adam optimizer for the main effects, second-order effects, and model tuning are all 0.0001. Model accuracy is evaluated with mean squared error. The obtained interpretable neural network model achieved 0.0001 mean squared error and 0.7346 R-squared on the testing set which indicates sufficient accuracy.

Figure 5(*b*) shows that the interaction between the variables x_5 and x_{10} has the highest impact on the output, accounting for 56.2% of the total variation. Following that comes another second-order effect between x_9 and x_{10} but with only 18.3% importance. All main effect contributions are less than 10% with x_{10}, x_6, x_5, and x_9 standing out. Significant interactions between x_5, x_{10} and x_9, x_{10} are mathematically valid with the consideration of the terms $2x_5x_{10}$ and $1/(|3x_9| + |3x_{10}|)$ in Eq. (25). Considering the sampling distributions, multiplication, and multiplicative inverse of absolute value sums mathematical operations are expected to have high impacts on the function output.

We also employed the Sobol sensitivity analysis method for the same problem to validate the results of the introduced data-driven GSA approach. The same dataset with 10,000 samples is used in the analysis. The most significant effects are observed as x_5, x_{10} the second-order effect with 0.3368 and x_9, x_{10} second--order effect with 0.2099 Sobol sensitivity indices. Following these come x_{10}, x_9, x_6, x_5, x_1, and x_2 interaction, x_3 and x_4 interaction, and finally, x_5 and x_6 interaction, all having negligible sensitivity indices compared to the two dominant second-order effects of x_5, x_{10} and x_9, x_{10} as presented in Fig. 5(*a*).

When Figs. 5(*a*) and 5(*b*) are examined, Sobol GSA results are consistent with the data-driven GSA results regarding the ranking of the important effects as x_5, x_{10} and x_9, x_{10} are identified as the major contributors in both methods. Minor differences are observed between the two methods for some of the less significant effects. The importance ranking and the strength of the effect differ for x_9, x_6, and x_5 in two methods. Besides, Sobol sensitivity analysis reveals second-order effects between x_3 and x_4, x_1 and x_2, x_5 and x_6 which are not detected by the data-driven GSA method. This is an expected result as the ANN model prunes insignificant main and second-order effects; additionally, second-order effects can further be discarded if none of the parent main effects are significant. Overall, the Sobol sensitivity analysis results validate the data-

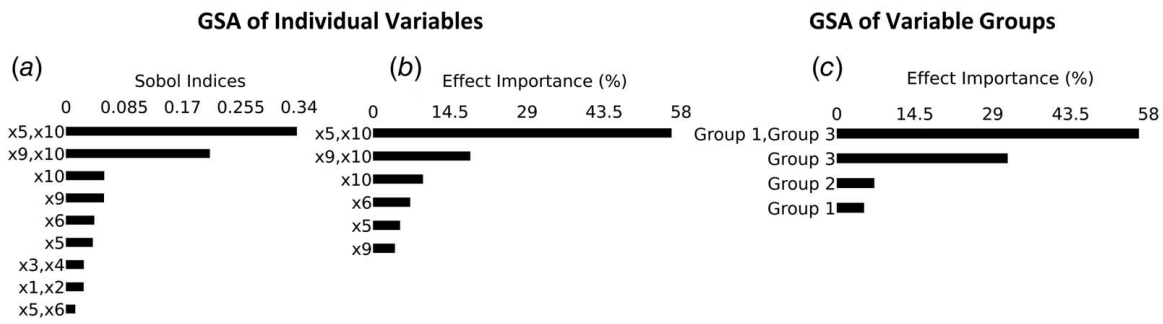


Fig. 5 Global sensitivity analysis results of the mathematical example (*a*) with Sobol's method for individual variables, (*b*) with interpretable neural networks for individual variables, and (c) with interpretable neural networks for variable groups

driven GSA method within the limits of previously defined modeling constraints for managing model complexity.

Regarding the computational efficiency of both methods, Sobol GSA takes a couple of seconds whereas data-driven GSA requires a couple of minutes for this problem (with a known analytical model) on one single-core CPU and 12.7 GB RAM. However, such an increase is not worth considering because when using Sobol's method in real applications, more computational time needs to be allocated to surrogate modeling than GSA itself before the Sobol GSA can even be applied. In addition, compared to the time used for data collection through either physical experiments or physics-based simulations, the time for GSA is often negligible. No matter which technique is used, the computational cost for GSA is expected to grow as the number of input variables increases. In general, an interpretable machine learning-based approach is better suited when the underlying physical relationship is complex and when the data are sufficient.

As for the dataset size, results for a dataset of 10,000 samples have been presented in this study. With the purpose of accurate comparison between Sobol's method and the proposed method, such a large dataset is employed in both analyses considering that Sobol's method uses a surrogate model which makes it affordable to work with large datasets. On the other hand, the proposed data-driven GSA approach can work with a large dataset provided by a surrogate model or be used as an end-to-end approach where a surrogate model is not necessary and sufficient data for an accurate ANN model is available from physical experiments/simulations. The effect of dataset size is summarized in Fig. 6 where the proposed approach is adopted for datasets ranging from 1000 samples to 10,000 samples. The mean squared error of all models is at acceptable levels 0.0021 being the highest error observed for the model trained on 1000 samples. R-squared value of all models similarly indicates accurate models with 0.84 being the lowest value for the model trained on 1000. Out of all possible main and second-order effects, all models identify the same effects as the top six effects which are x_5 and x_{10}, x_9 and x_{10}, x_{10}, x_6, x_5, x_9. The ranking of these effects shows variations for models trained on smaller datasets, yet becomes stable for datasets with 4000–5000 samples or larger in this problem with 10 input variables. The actual values of the effect importances follow similar proportions in all models, still showing variations. In brief, for models that successfully capture the physical relations between variables, the same important effects and importance ranking are anticipated for datasets with varying sample sizes. Considering the purpose of GSA, identification of the important effects and importance ranking has higher importance compared to the numerical values of the effect importance. It should also be noted that compared to the time used for data collection through either physical experiments or physics-based simulations, the time for GSA is often negligible.

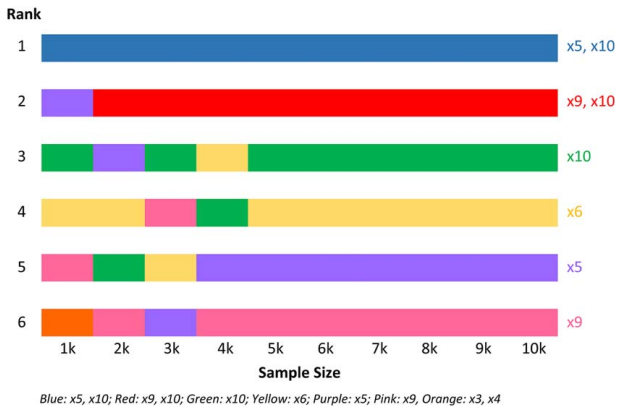


Fig. 6 Global sensitivity analysis results of the mathematical example for datasets with 1000 to 10,000 samples

4.1.3 Data-Driven Global Sensitivity Analysis of Variable Groups. The proposed approach allows extending the GSA results of individual variables to variable groups. Having sampled every variable from independent distributions, no correlations between the inputs exist. Thus, we can form variable groups within the input space using the subset decomposition method.

Equation (26) contains five terms associated with the function. Intuitively, three variables groups contained in the same mathematical operations can be defined such that $Group_1 = \{x_1, x_2, x_3, x_4, x_5\}$, $Group_2 = \{x_6, x_7, x_8\}$, $Group_3 = \{x_9, x_{10}\}$

$$f(\boldsymbol{x}) = \tanh(x_1x_2 + x_3x_4)\sqrt{|x_5|} + 0.3e^{x_5+x_6} + \log((x_6x_7x_8)^2 + 1) + 2x_5x_{10} + \frac{1}{|3x_9| + |3x_{10}|} \tag{26}$$

Figure 5(*c*) displays that the second-order effects between the variable groups outweigh the main effects. Variables from $Group_1$ and $Group_2$, similarly from $Group_1$ and $Group_3$ appear together within the same mathematical operation terms in Eq. (26). Thus, interactions are expected between $Group_1$, $Group_2$ and $Group_1$, $Group_3$. The second-order effect between $Group_1$ and $Group_3$ accounts for 56.2% importance while the main effects of $Group_1$ and $Group_3$ are 5.1% and 31.8%, respectively. Interestingly, the second-order effect significantly exceeds the main effects of both parent groups. This result sets an example for the benefit of analyzing the interplay between variable groups. It is crucial in this problem to analyze how these two groups influence each other and create a joint effect that is more critical than the parent groups.

4.2 Programmable Photonic Metasurface Design

4.2.1 Background. Photonic metasurfaces are artificially engineered structures that can support sophisticated light-matter interaction through subwavelength inclusions [31,32]. Advancements in the design and fabrication of photonic metasurfaces enable remarkable functionalities such as perfect absorption, super-resolution imaging, sensing, waveguiding, and invisibility cloak.

Programmable photonic metasurfaces are a special type of photonic metasurfaces that can transform between different functional states as a response to external stimuli. This characteristic enables programming these structures to switch certain properties under changing external stimuli for performing. To enable programmable photonic metasurfaces, a diverse array of physical mechanisms has been reported in the photonic communities, such as mechanical [33,34], thermal [35,36], electric [37–39], chemical [40,41], and light [42–44].

Light-based programmable photonic metasurfaces, which are the focus of this article, transform their functional state when stimulated with light and involve the simultaneous design of multiple systems, namely material, architecture, and stimulus. Stimulus, an input electromagnetic loading at a high level, deserves separate attention for modeling and design as it allows rich design freedom jointly formed by amplitude, phase, and polarization. In some prior work, the whole two-dimensional incident field was viewed as the stimulus to be designed. Ideally, material, architecture, and stimulus should be modeled and designed concurrently to ensure transparency and avoid suboptimality. Still, the common practice has been specifying the material (e.g., dielectric; metallic) and stimulus (e.g., a single frequency or a frequency band; polarization type) a priori and then only modeling/designing with respect to architecture. To this end, we analyze the complex interactions between these design entities and restate the design space with simple expressions using the data-driven GSA method for variable groups of material, architecture, and stimulus.

4.2.2 Design Problem. Figure 7(*a*) depicts a photonic metasurface which contains subwavelength structures and is responsive to light. When light is exerted on the metasurface, it is reflected and transmitted at certain amounts. The system response of interest for the design problem is this light transmission property. The

design problem involves architecture design, design of the subwavelength structures, and stimulus design. Architecture variable group contains three variables one being the unit cell type, a categorical variable; and two continuous variables for parametrizing the unit cell designs (Fig. 7(*d*)). As for the stimulus, two variables involve the excitation frequency of the light (Fig. 7(*b*)), a continuous variable, and polarization property of the light (Fig. 7(*c*)) which is a categorical variable.

The stimulus variable group includes frequency which is a continuous input variable in our model as shown in Fig. 7(*b*). Excitation frequency is inversely proportional to wavelength λ. The ratio between λ and a characteristic length scale of the system (e.g., periodicity Λ in a metasurface) primarily governs the light-matter interaction. Depending on the order of the ratio λ/Λ, the associated behavior of light-matter interaction tends to vary significantly in relation to different physical mechanisms.

The second input variable in the stimulus group is polarization which is illustrated in Fig. 7(*c*). Polarization is a property of transverse waves, and it characterizes the orientation of the field oscillations. An electromagnetic wave, an instance of transverse waves, contains electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{H}$, both of which have orthogonal directions to the wave propagation direction $\boldsymbol{k}$. Polarization is conventionally described by stating the electric field direction. In this article, we consider two types of polarizations: *x*-directional and *y*-directional linear polarization which are treated as categorical variables in the model. Provided that the wave propagation is given as $\boldsymbol{k} = (0, 0, k)$, each polarization is described by the notation of Johns vector as shown in Eqs. (27) and (28).

$$\begin{pmatrix} E_{0x}e^{i\varnothing_x} \\ E_{0y}e^{i\varnothing_y} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{27}$$

$$\begin{pmatrix} E_{0x}e^{i\varnothing_x} \\ E_{0y}e^{i\varnothing_y} \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -i \end{pmatrix} \tag{28}$$

Under the architecture variable group, we examine unit cell design as described in Fig. 7(*d*). Photonic metasurfaces contain periodic subwavelength features as the major building blocks in architectural design. The cross-sectional geometry of the building blocks can take free form without any restrictions. In this article, we are particularly interested in four canonical classes of unit cells reported in photonics communities. Two continuous unit cell design parameters create parametric variation within each unit cell class.

The system response of major interest is transmission that quantifies the energy transport from input to output in a two-port system as shown in Fig. 7(*b*). We aim to reveal the relationships between architecture-transmission, stimulus-transmission, and their interaction on transmission. The power transport between the ports is described by an S-parameter matrix whose individual components correspond to pairwise power transports. Given an *n*-port system with port *k* as the input port, the power transport S_{ij} from port *i* to port *j* of the electromagnetic system is computed with Eq. (29) [45] where $\boldsymbol{E}_c$ is the computed electric field that includes both excitation and scattered field, A_i is the face of port *i*, and † is the conjugate operator. This simulation can be viewed as a two-port network with the excitation port at the top face (Γ_1 in Fig. 7(*a*)) and the listener port at the bottom face (Γ_2 in Fig. 7(*a*)). From the *S*-parameter matrix, transmission is formulated with Eq. (30) where ω is the excitation angular frequency.

$$S_{ij} = \begin{cases} \dfrac{\int_{A_i}((\boldsymbol{E}_c - \boldsymbol{E}_i).\boldsymbol{E}_i^{\dagger})dA_i}{\int_{A_i}(\boldsymbol{E}_i.\boldsymbol{E}_i^{\dagger})dA_i} & i = k \\ \dfrac{\int_{A_i}(\boldsymbol{E}_c.\boldsymbol{E}_i^{\dagger})dA_i}{\int_{A_i}(\boldsymbol{E}_j.\boldsymbol{E}_i^{\dagger})dA_i} & \text{otherwise} \end{cases} \tag{29}$$

$$T(\omega) = |S_{21}(\omega)|^2 \tag{30}$$

4.2.3 Dataset Construction. The incident wave, stimulus in the design problem, can be viewed as an electromagnetic loading condition and is illuminated from the top face, propagating along the -*z*-direction. It is a plane wave specified by two input conditions: excitation frequency $f \in I_F = [30, 60]$ THz and two polarization types, *x*-directional and *y*-directional linear polarization.

As for the architecture, we consider four different geometric families. The cross section is extruded along the *z*-direction with the height $H = 1000$ nm. The periodicity Λ of the analysis domain is set as $\Lambda = 2800$ nm. Assumed to be lossless, the refractive index *n* of the metasurface and the SiO_2 substrate is set as $n = 5$ and $n = 1.45$, respectively. All the lateral faces Γ are subject to periodic boundary conditions. This setting effectively mimics the periodical tessellation of identical, infinitely many building blocks on the *xy*-plane.

The full-wave analysis is conducted by the RF module of COMSOL Multiphysics® [45]. The simulations for the data generation process formulate and solve the differential form of transmission spectra, system response in the design problem, according to Eq. (29) together with the initial and boundary conditions. The equations are solved using the finite element method with numerically stable edge element discretization.

The mapping of interest is constructed as $U \times G \times P \times f \rightarrow T$ where $U = \{c_j | j = 1, 2, 3, 4\}$ is the set of unit cell types, $g =$

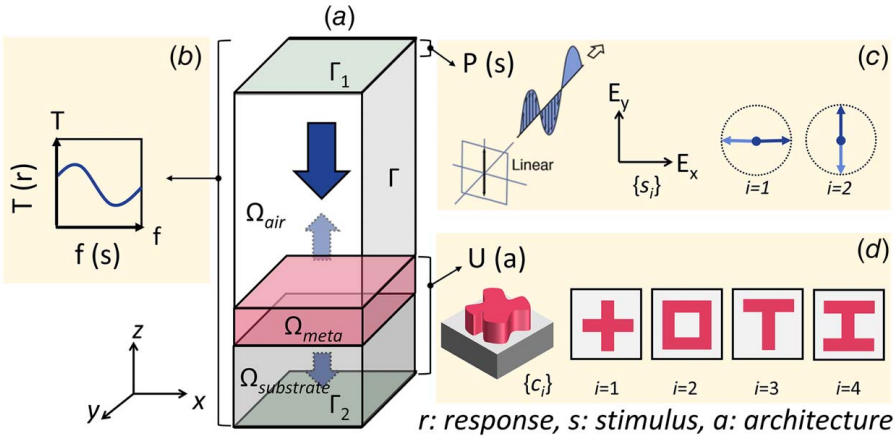


Fig. 7 Schematic of the programmable metasurface design problem (*a*) wave analysis, (*b*) transmission as response and frequency as stimulus, (*c*) polarization as stimulus, and (*d*) unit cell type as architecture variables

$\{g_j|j=1,2\} \in G \subset \mathbb{N}^2$ is the continuous vector that specifies parametric variation within G with respect to a given unit cell type c_j, $P=\{s_i|i=1,2\}$ is the set of polarization types, $f \in I_F = [30, 60]$ THz is the excitation frequency, and $T \in [0, 1]$ is the transmission. To generate a dataset D, 20 space-filling designs are sampled from G using the Latin hypercube sampling method. The frequency band I_F is discretized with a spacing of $\Delta f = 1$ THz. As a result, D includes $|D| = |U| \times |G| \times |P| \times (|I_F|/\Delta f) = 4960$ observations.

4.2.4 Data-Driven Global Sensitivity Analysis of Individual Variables. We start with training an interpretable ANN model for this problem [30]. The testing to training datasets are obtained by randomly splitting the dataset with a 0.8/0.2 ratio. Main effects and the most meaningful first 20 second-order effects are considered. Each subnetwork has five ReLU hidden layers with 40 nodes per layer. The batch size is set to 512 while the maximum number of epochs is 1000, and the learning rates of the Adam optimizer are 0.0001 for the main effects, interaction effects, and model tuning. Model accuracy is evaluated with mean squared error. The obtained interpretable neural network model achieved 0.0022 mean squared error and 0.9793 R-squared on the testing set.

Figure 8(*a*) indicates that excitation frequency steps up with 40.3% of the total effects on transmission, followed by the second-order effect between unit cell type and frequency with 37.1%, and finally the main effect of unit cell type with 21.2% importance. Polarization on its own and its interactions with other variables exhibit less significant impact compared to the others.

4.2.5 Data-Driven Global Sensitivity Analysis of Variable Groups. We examine the groups of input variables to further analyze the importance of variable group main and second-order effects instead of individual consideration of each input variable. Since we sampled every variable from independent distributions, no correlations between the inputs exist. Thus, variable groups can be defined using the subset decomposition method.

The design problem naturally contains variables related to architecture and stimulus such that Architecture $=\{U, G\}$, Stimuli $=\{P, f\}$. Here, we are interested in inferring the importance of these groups as well as any interactions between them for summarizing the complex design space and inferring guidelines for the design methodology.

Figure 8(*b*) indicates stimulus as being the most important variable group for explaining the metasurface response with 41.4% importance. The other variable group, architecture, accounts for 21.2%. Designing architecture on its own or similarly designing stimulus in isolation would be an inadequate approach for this case as the interaction between stimulus and architecture has a substantial effect on the response with 37.4%. In the following section, we further analyze the architecture–stimulus relation and their combined influence on the system response.

4.2.6 Discussion. Figure 8(*b*) illustrates the importance of designing architecture and stimulus together as their combined effect accounts for 37.4% of the total variance in the model response. For this purpose, impacts of the input variables are visually assessed with PDP which was introduced in Sec. 2. The main effects of polarization, frequency, and unit cell type are illustrated in Figs. 9(*a*)–9(*c*). The critical contribution of architecture and stimulus is then further studied with the second-order effects of polarization and unit cell type in Fig. 9(*d*), and frequency and unit cell type in Fig. 9(*e*).

Figure 9(*a*) shows that the two polarization types covered in this article, s_1 and s_2, have similar effects on the system response. Hence, the two classes can be used interchangeably.

Figure 9(*b*) indicates that when the frequency is below 47 THz, transmission reaches the maximum with a value of 1 and settles to 0.5 when the frequency is above 52 THz. Between these two frequencies, for [47, 52] THz, a steep decrease occurs.

Figure 9(*c*) presents the effect of unit cell type. c_1 and c_2 display identical effects, maximizing the transmission, and a similar situation is observed for c_3 and c_4 resulting in a transmission of 0.7. Therefore, if transmission is of primary interest for design, $c_1 - c_2$ and $c_3 - c_4$ can be interchangeably used without significant changes in the system response.

Figure 9(*d*) shows the second-order effect of the polarization and unit cell type and does not offer any unexpected insights in addition to Figs. 9(*a*) and 9(*c*). Again, the polarization type does not have any meaningful contribution while unit cell types $c_1 - c_2$ and $c_3 - c_4$ create similar effects.

When it comes to frequency and unit cell type interaction, insightful observations appear in Fig. 9(*e*). When frequency is set below 52 THz, the transmission response of the photonic metasurface becomes maximum (unity). This system response is observed regardless of the unit cell type; no matter what architecture design is preferred, the maximum system response is attained. When the frequency is set above 52 THz, the system response becomes either maximum or minimum depending on the unit cell type. To elaborate, c_1 and c_2 maximize the transmission while c_3 and c_4 minimize it, revealing that for frequencies higher than 52 THz,

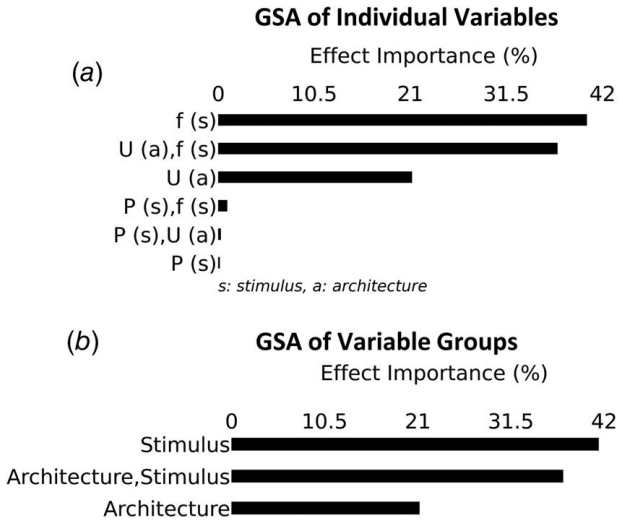


Fig. 8 (*a*) Data-driven global sensitivity analysis results of the programmable photonic metasurface for (*b*) individual variables for variable groups

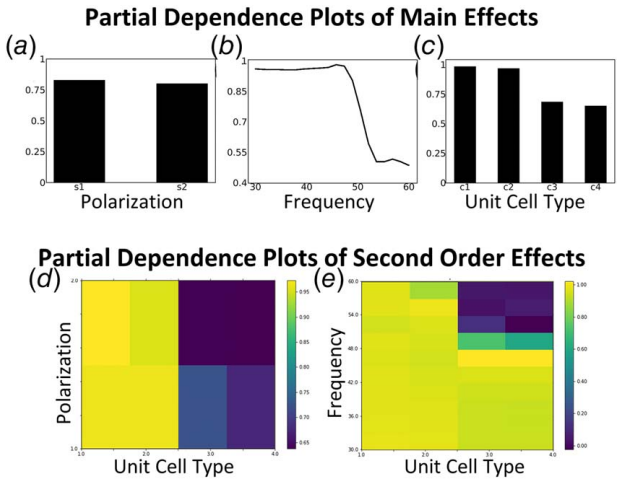


Fig. 9 Programmable metasurface design problem partial dependence plots of (*a*) polarization, (*b*) frequency, (c) unit cell type, (*d*) polarization and unit cell type, and (*e*) frequency and unit cell type

Table 1 Design guidelines for a programmable photonic metasurface with two functional states

Maximum response: $T = 1$		Minimum response: $T = 0$	
Stimulus	Architecture	Stimulus	Architecture
$f > 52$ THz	c_1, c_2	$f < 52$ THz	c_3, c_4
$f < 52$ THz	c_1, c_2, c_3, c_4		

architecture plays a critical role. This result validates the necessity of simultaneous consideration of stimulus and architecture since the interaction between them brings about a unique effect on the model response.

Table 1 summarizes the conclusions we derive for achieving two functional states with this programmable photonic metasurface. To set the PMS to maximum transmission, two alternative configurations are available and as for the other functional state when the transmission is minimum, just one configuration is convenient. $f <$ 52 THz allows achieving both states, thus is preferable when using two different unit cells is inexpensive. Similarly, c_1 or c_2 also can result in both states and can be a better option when using two different excitation frequencies is inexpensive.

The proposed method suggests that it is sufficient to employ one frequency and two unit cells or two frequencies and one unit cell for designing a PMS with two functional states. With this, we reduce a complex problem with a large input space containing many parameters to just a few design inputs.

The interpretations obtained from the data-driven GSA for variable groups serve as an input for managing the design complexity and deriving design guidelines for PMS development by providing answers to the following:

(1) Is it possible to obtain a set of functional states with just one material or is a spatially varying combination of multiple materials necessary?
(2) Is it possible to obtain the target performance with just designing the architecture variables or is the concurrent design of architecture and stimuli required?
(3) Which architectural structures provide a highly diversified set of functional states when programmed with the stimulus?

5 Conclusion

In this article, we introduce a data-driven approach for performing GSA based on interpretable neural networks and further extend the analysis results for variable groups. The proposed method performs the FANOVA decomposition of a machine learning model for partitioning the output variance into terms associated with the inputs, then identifies the contributions of the variable groups on the model response. We demonstrate that the implementation of the proposed method is valuable for understanding complex physical interactions in engineering design for which system models cannot be explicitly derived due to system complexity. Finally, we present a use case and potential benefits of adopting a data-driven approach for conducting GSA in the real engineering design of PMS.

The introduced data-driven methodology provides substantial benefits to the global sensitivity analysis of variable groups. First, it is an end-to-end approach that eliminates the necessity of developing analytical or surrogate models, which can become expensive to work with or difficult to obtain. Being a data-driven approach, the method can successfully manage large datasets, high-dimensional problems with numerous variables, mixed input spaces containing both categorical and continuous variables, and highly nonlinear system behavior. Moreover, global and local sensitivity analysis in both regression and classification problems can be performed and visually interpreted. The comparative study of datasets with various sample sizes revealed that the proposed approach is successful in identifying the most important variable effects even in small datasets. Increasing the dataset size is suggested when data collection costs are manageable as it improves the predictive accuracy of the model as well as the robustness of the variable importance ranking. The PMS application presented in this article shows that the approach is valuable for managing the design space complexity when working with large input spaces to extract the most meaningful design entities. These conclusions can then drive the derivation of design guidelines for PMS development.

We further identify some future efforts with potential benefits for advancing the introduced approach. First, the interpretable neural network model used in this article only covers the main effects and second-order effects. Adjusting the model architecture for considering higher-order effects can generate more accurate models. Similarly, the interpretable neural network model is suitable for single-output problems. Considering that many engineering problems involve multiple outputs, modifying the model structure for multi-output problems is a promising effort. Finally, we plan on analyzing more complex PMS problems where there are more input variables with interactions between all three groups of material, architecture, and stimulus variables to ensure the transferability of the previously obtained design rules [46].

Funding Data

- The National Science Foundation (NSF) Boosting Research Ideas for Transformative and Equitable Advances in Engineering (BRITE) fellow program (award number 2227641).
- The Cyberinfrastructure for Sustained Scientific Innovation (CSSI) program (award number 1835782).
- Army Research Laboratory (cooperative agreement number W911NF-22-0121).

Conflict of Interest

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Iooss, B., and Lemaître, P., 2015, "A Review on Global Sensitivity Analysis Methods," *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications: Operations Research/Computer Science Interfaces Series*, G. Dellino, and C. Meloni, eds., Springer US, Boston, MA, pp. 101–122.
[2] Chatterjee, S., and Hadi, A. S., 2009, *Sensitivity Analysis in Linear Regression*, John Wiley and Sons, Hoboken, NJ.
[3] Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S., 2010, "Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity Index," Comput. Phys. Commun., **181**(2), pp. 259–270.
[4] Christopher Frey, H., and Patil, S. R., 2002, "Identification and Review of Sensitivity Analysis Methods," Risk Anal., **22**(3), pp. 553–578.
[5] Miller, T., 2019, "Explanation in Artificial Intelligence: Insights From the Social Sciences," Artif. Intell., **267**, pp. 1–38.
[6] Chen, W., Allen, J. K., Tsui, K.-L., and Mistree, F., 1996, "A Procedure for Robust Design: Minimizing Variations Caused by Noise Factors and Control Factors," ASME J. Mech. Des., **118**(4), pp. 478–485.
[7] Kim, Y., Yuk, H., Zhao, R., Chester, S. A., and Zhao, X., 2018, "Printing Ferromagnetic Domains for Untethered Fast-Transforming Soft Materials," Nature, **558**(7709), pp. 274–279.
[8] Yang, Z., Zhang, A., and Sudjianto, A., 2021, "GAMI-Net: An Explainable Neural Network Based on Generalized Additive Models With Structured Interactions," Pattern Recognit., **120**, p. 108192.

[9] Chen, W., Jin, R., and Sudjianto, A., 2004, "Analytical Variance-Based Global Sensitivity Analysis in Simulation-Based Design Under Uncertainty," ASME J. Mech. Des., **127**(5), pp. 875–886.
[10] Molnar, C., 2020, *Interpretable Machine Learning*, Lulu.com, Research Triangle, NC.
[11] Friedman, J. H., 2001, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Stat., **29**(5), pp. 1189–1232.
[12] Apley, D. W., and Zhu, J., 2020, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," J. R. Stat. Soc. B, **82**(4), pp. 1059–1086.
[13] Breiman, L., 1996, "Bagging Predictors," Mach. Learn., **24**(2), pp. 123–140.
[14] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E., 2015, "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," J. Comput. Graph. Stat., **24**(1), pp. 44–65.
[15] Ribeiro, M. T., Singh, S., and Guestrin, C., 2016, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, ACM, pp. 1135–1144.
[16] Lundberg, S. M., and Lee, S. I., 2017, "A Unified Approach to Interpreting Model Predictions," Adv. Neural Infor. Proc. Syst., **30**.
[17] Marchese Robinson, R. L., Palczewska, A., Palczewski, J., and Kidley, N., 2017, "Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets," J. Chem. Inf. Model., **57**(8), pp. 1773–1792.
[18] Baryannis, G., Dani, S., and Antoniou, G., 2019, "Predicting Supply Chain Risks Using Machine Learning: The Trade-off Between Performance and Interpretability," Future Gener. Comput. Syst., **101**, pp. 993–1004.
[19] Dunnington, D. W., Trueman, B. F., Raseman, W. J., Anderson, L. E., and Gagnon, G. A., 2021, "Comparing the Predictive Performance, Interpretability, and Accessibility of Machine Learning and Physically Based Models for Water Treatment," ACS ES&T Eng., **1**(3), pp. 348–356.
[20] Johansson, U., Sönströd, C., Norinder, U., and Boström, H., 2011, "Trade-Off Between Accuracy and Interpretability for Predictive in Silico Modeling," Future Med. Chem., **3**(6), pp. 647–663.
[21] Özesmi, S. L., and Özesmi, U., 1999, "An Artificial Neural Network Approach to Spatial Habitat Modelling With Interspecific Interaction," Ecol. Modell., **116**(1), pp. 15–31.
[22] David, G. G., 1991, "Interpreting Neural-Network Connection Weights," AI Expert, **6**(4), pp. 46–51.
[23] Scardi, M., and Harding, L. W., 1999, "Developing an Empirical Model of Phytoplankton Primary Production: A Neural Network Case Study," Ecol. Modell., **120**(2), pp. 213–223.
[24] Dimopoulos, Y., Bourret, P., and Lek, S., 1995, "Use of Some Sensitivity Criteria for Choosing Networks With Good Generalization Ability," Neural Process. Lett., **2**(6), pp. 1–4.
[25] Gevrey, M., Dimopoulos, I., and Lek, S., 2003, "Review and Comparison of Methods to Study the Contribution of Variables in Artificial Neural Network Models." Ecol. Modell., **160**(3), pp. 249–264.
[26] Sobol′, I. M., 2001, "Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates," Math. Comput. Simul., **55**(1), pp. 271–280.
[27] Campolongo, F., Cariboni, J., and Saltelli, A., 2007, "An Effective Screening Design for Sensitivity Analysis of Large Models," Environ. Modell. Softw., **22**(10), pp. 1509–1518.
[28] Sobol, I. M., and Kucherenko, S., 2010, "Derivative Based Global Sensitivity Measures," Procedia Soc. Behav. Sci., **2**(6), pp. 7745–7746.
[29] Hooker, G., 2007, "Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables," J. Comput. Graph. Stat., **16**(3), pp. 709–732.
[30] Python, n.d., "PiML 0.5.0.Post1: A Low-Code Interpretable Machine Learning Toolbox in Python," https://github.com/SelfExplainML/PiML-Toolbox, Accessed September 7, 2023.
[31] Bukhari, S. S., Vardaxoglou, J. (., and Whittow, W., 2019, "A Metasurfaces Review: Definitions and Applications," Appl. Sci., **9**(13), p. 2727.
[32] Chen, H.-T., Taylor, A. J., and Yu, N., 2016, "A Review of Metasurfaces: Physics and Applications," Rep. Prog. Phys., **79**(7), p. 076401.
[33] Specht, M., Berwind, M., and Eberl, C., 2021, "Adaptive Wettability of a Programmable Metasurface," Adv. Eng. Mater., **23**(2), p. 2001037.
[34] Liu, S., Zhang, L., Bai, G. D., and Cui, T. J., 2019, "Flexible Controls of Broadband Electromagnetic Wavefronts With a Mechanically Programmable Metamaterial," Sci. Rep., **9**(1), p. 1809.
[35] Lor, C., Phon, R., Lee, M., and Lim, S., 2022, "Multi-Functional Thermal-Mechanical Anisotropic Metasurface With Shape Memory Alloy Actuators," Mater. Des., **216**, p. 110569.
[36] Yin, H., Liang, Q., Duan, Y., Fan, J., and Li, Z., 2022, "3D Printing of a Thermally Programmable Conformal Metasurface," Adv. Mater. Technol., **7**(7), p. 2101479.
[37] Shirmanesh, G. K., Sokhoyan, R., Wu, P. C., and Atwater, H. A., 2020, "Electro-Optically Tunable Multifunctional Metasurfaces," ACS Nano, **14**(6), pp. 6912–6920.
[38] Wan, X., Qi, M. Q., Chen, T. Y., and Cui, T. J., 2016, "Field-Programmable Beam Reconfiguring Based on Digitally-Controlled Coding Metasurface," Sci. Rep., **6**(1), p. 20663.
[39] Fu, X., Shi, L., Yang, J., Fu, Y., Liu, C., Wu, J. W., Yang, F., Bao, L., and Cui, T. J., 2022, "Flexible Terahertz Beam Manipulations Based on Liquid-Crystal-Integrated Programmable Metasurfaces," ACS Appl. Mater. Interfaces, **14**(19), pp. 22287–22294.
[40] Dong, S., Zhang, K., Yu, Z., and Fan, J. A., 2016, "Electrochemically Programmable Plasmonic Antennas," ACS Nano, **10**(7), pp. 6716–6724.
[41] Li, J., Chen, Y., Hu, Y., Duan, H., and Liu, N., 2020, "Magnesium-Based Metasurfaces for Dual-Function Switching Between Dynamic Holography and Dynamic Color Display," ACS Nano, **14**(7), pp. 7892–7898.
[42] Kao, T. S., Rogers, E. T. F., Ou, J. Y., and Zheludev, N. I., 2012, "'Digitally' Addressable Focusing of Light Into a Subwavelength Hot Spot," Nano Lett., **12**(6), pp. 2728–2731.
[43] Buijs, R. D., Wolterink, T. A. W., Gerini, G., Verhagen, E., and Femius Koenderink, A., 2021, "Programming Metasurface Near-Fields for Nano-Optical Sensing," Adv. Opt. Mater., **9**(15), p. 2100435.
[44] Lee, D., Jiang, S., Balogun, O., and Chen, W., 2022, "Dynamic Control of Plasmonic Localization by Inverse Optimization of Spatial Phase Modulation," ACS Photonics, **9**(2), pp. 351–359.
[45] COMSOL Multiphysics®, n.d., www.comsol.com, COMSOL AB, Stockholm, Sweden.
[46] Dolar, T., Lee, D., and Chen, W., 2023, "Interpretable Neural Network Analyses for Understanding Complex Physical Interactions in Engineering Design," International Design Engineering Technical Conferences, Boston, MA, Aug. 20–23, American Society of Mechanical Engineers, Vol. 87301, p. V03AT03A021.