Data Centric Nanocomposites Design via Mixed-Variable Bayesian Optimization

Journal:	*Molecular Systems Design & Engineering*
Manuscript ID	ME-ART-06-2020-000079.R1
Article Type:	Paper
Date Submitted by the Author:	03-Aug-2020

Complete List of Authors:	Iyer, Akshay; Northwestern University, Zhang, Yichi; Northwestern University, Mechanical Engineering Prasad, Aditya; Rensselaer Polytechnic Institute, Material Science and Engineering Gupta, Praveen; Rensselaer Polytechnic Institute Tao, Siyu; Northwestern University Wang, Yixing; Northwestern University, Mechanical Engineering Prabhune, Prajakta; Duke University Schadler, Linda; University of Vermont Brinson, L.; Duke University, Mechanical Engineering and Materials Science Chen, Wei; Northwestern University, Mechanical Engineering

Design, System, Application

While polymer nanocomposites display an unprecedented combination of mechanical and electrical properties, tailoring them to meet application specific requirements remains a challenging task, owing to the vast, mixed variable design space that includes composition (i.e. choice of polymer, nanoparticle & surface modification) and microstructures of nanocomposite material. Modelling properties of interphase region introduces additional complexity to the design process and requires computationally expensive simulations.

This article demonstrates that a data centric design framework, where each step of design process is guided by experimental and/or simulated data, can overcome these challenges. Using design of nanocomposites for electrical insulation as an exemplar, we describe the integration of experimental data with sophisticated computational simulations for microstructure characterization, interphase modelling, and structure-property prediction. A novel Latent Variable Gaussian Process (LVGP) approach enables mixed variable Bayesian Optimization for concurrent composition and microstructure optimization to expedite the search for Pareto designs under multiple performance criteria. While discussions are centered on nanocomposites, the concepts of data centric design, mixed variable Bayesian Optimization and multicriteria design are ubiquitous and immediately applicable to other material systems.

Data Centric Nanocomposites Design via Mixed-Variable Bayesian Optimization

Akshay Iyer[1], Yichi Zhang[1], Aditya Prasad[2], Praveen Gupta[2], Siyu Tao[1], Yixing Wang[1],Prajakta Prabhune[4], Linda Schadler[3], L Catherine Brinson[4], Wei Chen[1*]

[1]Department of Mechanical Engineering, Northwestern University, USA

[2] Department of Material Science & Engineering Rensselaer Polytechnic Institute, USA

[3] Department of Engineering & Mathematical Sciences, The University of Vermont, USA

[4] Department of Mechanical Engineering & Material Science, Duke University, USA

ABSTRACT

With an unprecedented combination of mechanical and electrical properties, polymer nanocomposites have the potential to be widely used across multiple industries. Tailoring nanocomposites to meet application specific requirements remains a challenging task, owing to the vast, mixed-variable design space that includes composition (i.e. choice of polymer, nanoparticle, and surface modification) and microstructures (i.e. dispersion and geometric arrangement of particles) of the nanocomposite material. Modeling properties of interphase, the region surrounding a nanoparticle, introduces additional complexity to the design process and requires computationally expensive simulations. As a result, previous attempts at designing polymer nanocomposites have focused on finding the optimal microstructure for only a fixed combination of constituents. In this article, we propose a data centric design framework to concurrently identify optimal composition and microstructure using mixed-variable Bayesian Optimization. This framework integrates experimental data with state-of-the-art techniques in interphase modeling, microstructure characterization & reconstructions and machine learning. Latent Variable Gaussian Processes (LVGPs) quantifies the lack-of-data uncertainty over the mixed-variable design space that consists of qualitative and quantitative material design variables. The design of electrically insulating nanocomposites is cast as a multicriteria optimization problem with the goal of maximizing dielectric breakdown strength while minimizing dielectric permittivity and dielectric loss. Within tens of simulations, our method identifies a diverse set of

* Corresponding Author Email: weichen@northwestern.edu

designs on the Pareto frontier indicating the tradeoff between dielectric properties. These findings project data centric design, effectively integrating experimental data with simulations for Bayesian Optimization, as an effective approach for design of engineered material systems.

1. INTRODUCTION

The launch of the Material Genome Initiative (MGI) [1] has revolutionized the way advanced material systems are designed with targeted performance. MGI strives to elucidate the Processing-Structure-Property (PSP) relationships [2] for material design. A holistic design strategy for bi-directional traversal of PSP relationships requires us to address some key issues – cost effective processing techniques, microstructure representation and reconstruction, dimensionality reduction and tractable optimization techniques, to name a few. In the field of polymer nanocomposites, goal-oriented design has proven to be a difficult task due to several reasons.

First, limited understanding of complex polymer(matrix)-nanoparticle(filler) interactions and their influence on properties hinders the selection of the optimal combination from the vast space possible combinations. While finite element analysis (FEA) models have been developed to simulate structure-property relationships for polymer nanocomposites [3-5], modeling interphase behavior remains a prominent challenge. Researchers have investigated interphase behaviors and their origin both analytically and experimentally [5-7]. Recent experiments have demonstrated that the local polymer properties significantly change near the polymer surface via measurement of properties in model nanocomposites [7, 8]. While direct measurement of interphase properties in nanocomposites is challenging experimentally, one method to calculate the interphase properties is to inversely tune the parameters in micro-scale model constitutive equations or finite elements analysis using the bulk composite properties [3, 9-11]. However, this tuning procedure is very time-consuming given the complexity of experimental data and the simulation cost of FEA.

Second, the high dimensionality of nanocomposite microstructure requires specialized techniques for characterization of micrographs with reduced dimensionality and establish its relationship with processing conditions and properties. To this end, computational Microstructure Characterization and Reconstruction (MCR) [12] techniques provide a quantitative representation of microstructures and the ability to reconstruct realizations with desired features. Among the existing methods, Physical Descriptors [13, 14] and Spectral Density Function (SDF) [15-18] have

been widely adopted for design of material systems due to their physically meaningful characterization, relative ease of reconstruction and low dimensional representation. The selection of MCR method for a material system and ascertaining associated parameters is accomplished by analyzing the micrographs obtained from different processing conditions.

Third, calibration of interphase parameters and selection of MCR technique requires a database, where each nanocomposite sample is labelled by processing conditions, microstructure, and properties. NanoMine [19, 20] - a online database with built-in data curation capabilities provides access to several nanocomposites reported in the literature. However, articles seldom report all the aforementioned labels which hinders the development of PSP relationships necessary for targeted design of nanocomposites.

Fourth, the high computational cost of physics-based property evaluation methods prohibits their direct usage in the iterative design process that could require hundreds of property evaluations. To alleviate this problem, Bayesian Optimization (BO) [21, 22] has emerged as a viable proposition in material design [23-25]. However, these applications of BO involve only quantitative design variables in the form of descriptors (aka features) known to influence material properties; while mixed-variable problems containing both qualitative and quantitative variables is common in material design. Choice of constituents in any material system can be treated as qualitative variables, while microstructure descriptors, processing, and operating parameters (temperature, RPM, wavelength etc.) are quantitative variables. For example, nanocomposite design involves concurrent optimization of qualitative (choice of polymer, nanoparticle, surface modification) and quantitative (microstructure descriptors) variables. The Latent Variable Gaussian Process (LVGP) [26] provides an intuitive way to predict material properties from mixed-variable inputs and improves the performance of single criterion BO as compared to existing GP methods [27]. However, materials design requires mixed-variable multicriteria BO since suitability for commercial application relies heavily on multiple criteria.

These factors hinder the establishment of a comprehensive methodology to fully incorporate processing, structure, and property information for nanocomposite materials into the design process. Combinations of experimental, theoretical, and simulated investigations [28-32] have improved our understanding of the influence of materials and processing conditions on nanocomposite morphology and properties. These studies are typically guided by researcher's knowledge and intuition. In recent years, there has been a push toward the "fourth paradigm" of

science [33] which seeks to leverage the increasing data availability to develop tools that can effectively extract knowledge to guide a data-driven search of optimal materials. However, previous attempts at data-driven nanocomposite design have been limited to design of microstructure for a prespecified combination of polymer, nanoparticle and surface modification [34, 35].

This article presents a data-centric design framework and the associated techniques to leverage existing data for multicriteria nanocomposite design. The framework is flexible to incorporate data generated by experiments as well as simulations or machine learning to overcome existing challenges in establishing structure-property relationships. Nanocomposite design is cast as a mixed-variable optimization problem to concurrently identify optimal composition and microstructure. Central to the design strategy is integration of LVGP, which enables mixed-variable machine learning and uncertainty quantification, with multicriteria BO to navigate complex, non-linear design space and identify a diverse Pareto frontier. While discussions on data and modeling tools are centered on polymer nanocomposites, the concept of data centric design is generic and applicable to any material system.

2. DATA-CENTRIC NANOCOMPOSITE DESIGN FRAMEWORK

Despite their attractive mechanical and electrical properties, commercial application of polymer nanocomposites is plagued by a lack of goal-oriented design methodology. In this context, we present the data-centric design framework, guided by the philosophy that integrating curated databases with physics-based simulations and machine learning expedites nanocomposite design.

Fig.1 depicts the mixed-variable BO framework exemplified by the design of insulating materials, indicating the various modules involved and information flow between them. The framework is initiated from a materials database (**Module 1**) comprising nanocomposite samples with varying compositions, corresponding microstructures and measurement of properties such as dielectric loss. ***Composition*** is defined by the choices of polymer, nanoparticle and surface modification. ***Microstructure*** descriptors influenced by composition and processing conditions,

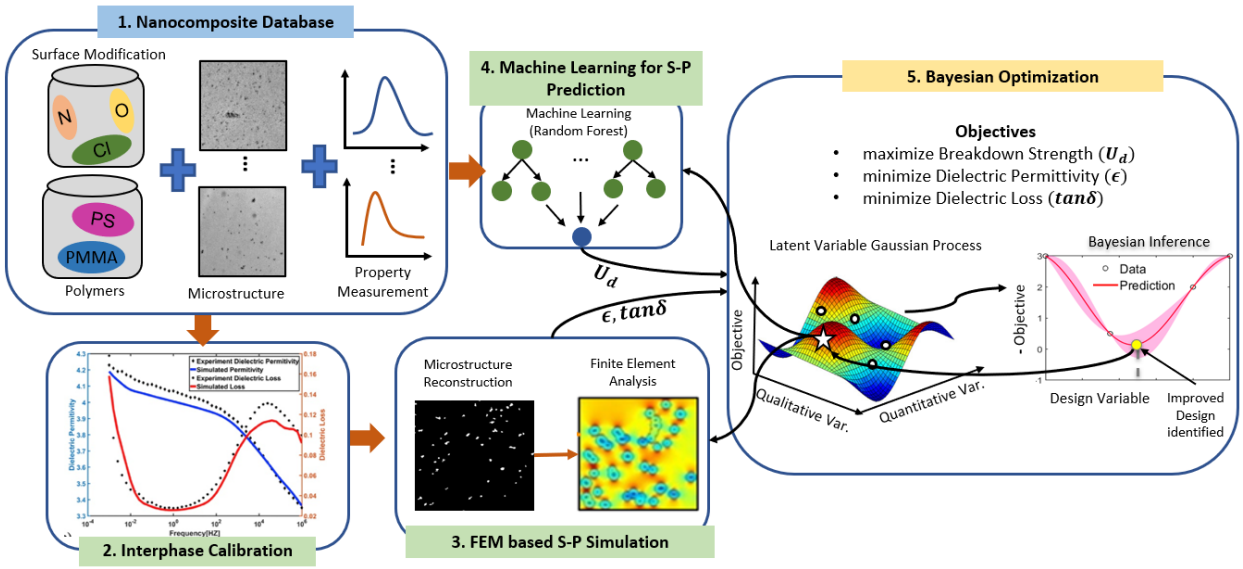

Figure 1: Data centric design framework for polymer nanocomposites

e.g., nanoparticle dispersion, are quantified from micrographs using the MCR techniques. The identified range of microstructure descriptors will be used as bounds in the design process.

The database also contains experimental measurements of nanocomposite properties, which can be used to calibrate simulation models (**Module 2**) and train machine learning models for situations where finite element simulations (**Module 3**) are too expensive or simulation models are premature (**Module 4**). For example, experimental measurements of bulk nanocomposites data are used for calibrating the nanoparticle-polymer interphase parameters necessary to accurately predict properties via FEA. With bounds for design variables identified and models to predict dielectric properties, BO (**Module 5**) expedites the search for high-performing nanocomposites designs. The steps included in the BO procedure can be summarized as follows:

I. A machine learning model is trained on existing data to predict material properties of interest from design variables and quantify prediction uncertainty

II. An acquisition function uses the predictions and associated uncertainties to select the design that promises the largest improvement in properties.

III. Properties of the selected design are evaluated and added to the dataset.

This process is repeated for a prespecified number of iterations or until a global optimum (for single criterion design) / Pareto front (for multicriteria design) is identified. While GP are frequently used in BO, existing GP models were developed for quantitative variables and the associated correlation functions cannot accommodate qualitative inputs. We overcome this

limitation by leveraging the recently developed LVGP[†] approach [26, 27] which implicitly converts qualitative variables to continuous latent variables for evaluating correlations. Since functional materials must satisfy multiple performance criteria, we extend the LVGP based BO for multicriteria optimizations.

In this article, we demonstrate the data-centric design process for electrically insulating polymer nanocomposites, with potential application in high voltage rotating machines [36]. Three major electrical properties to be optimized are breakdown strength, dielectric permittivity and dielectric loss. Breakdown strength (U_d) is the minimum voltage at which current flows through an insulating material. Dielectric permittivity (ϵ) characterizes the degree of electrical polarization experienced by the material and dielectric loss ($tan\delta$) is related to the amount of heat generated under an alternating electric field. High U_d, low ϵ and low $tan\delta$ are ideal but tradeoffs between U_d vs ϵ and ϵ vs $tan\delta$ have been observed [37, 38].

For the design of insulating materials, these properties are known to be influenced by composition (choice of filler, polymer, surface modification) and nanoparticle dispersion. We consider nanocomposites with two types of polymers - polystyrene (PS) and polymethylmethacrylate (PMMA) containing silica nanoparticles with three choices of surface modifications– Chloro-, Amino- and Octyl-silanes. Nanoparticle dispersion is quantified from Transmission Electron Microscopy (TEM) images using the Spectral Density Function (SDF) [15-18]. Dielectric permittivity ϵ and loss $tan\delta$ are evaluated using FEA, where interphase properties are characterized by a shift in the nanocomposite properties w.r.t pure polymer properties and obtained by calibration (**Module 2**) based on the bulk properties from experiments. In **Module 3** SDF based microstructure reconstruction [39] is used to generate 2D Representative Volume Elements (RVEs) with desired filler area fraction and dispersion for FEA. **Module 4** is an empirical machine learning model employing Random Forrest technique [40] which is trained on experimental data present in nanocomposite database to predict the breakdown strength U_d as a function of both qualitative and quantitative material design variables.

In **Module 5**, the mixed-variable BO problem is performed by leveraging the built-in uncertainty quantification of LVGP models for performing single and multicriteria optimization using the expected improvement [41] and expected maximin improvement [42] acquisition

[†] An implementation of LVGP in R programming language is available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=LVGP.

functions respectively. At each iteration, the LVGP model is updated with a new design whose dielectric properties are evaluated using Modules 3 and 4.

The design framework presented here has two significant benefits. First, its modularity allows for selection, replacement, and customization of methods within each module without affecting the rest of the framework. For example, the machine learning model used for U_d can be replaced by a physics-based simulation model in the future. The microstructure characterization & reconstruction method can be selected based on the nature (nanoparticle or nanotube) of the filler. Second, diverse applications can be explored using the same framework by modifying the objectives. For example, we can design nanodielectrics by maximizing ϵ and minimizing $U_d, tan\delta$ in Module 5 without modifying the rest of the framework.

3. IMPLEMENTING DATA CENTRIC DESIGN FRAMEWORK

In the following subsections, we describe the techniques that are used to support the implementation of the proposed materials design framework, using the design of insulating polymer composites as an example.

3.1 Nanocomposite Database Preparation (Module 1)

A database comprising nanocomposite samples labelled by their composition, processing conditions, microstructures and dielectric properties is essential for identifying design variables and developing the structure-property relations. For design of insulating nanocomposites, we developed a database of samples with varied composition and dispersions.

Silica nanoparticles (diameter 14 nm) in methyl ethyl ketone were procured from Nissan Inc. The surface of the nanoparticles was modified using three monofunctional silane coupling agents: aminopropyledimethylethoxysilane (Amino), chloropropyledimethylethoxysilane (Chloro) and octyldimethylmethoxysilane (Octyl), from Gelest Inc. Polystyrene (PS) from Goodfellow Corporation and polymethylmethacrylate (PMMA) from Scientific Polymer Products Incorporated is used as the polymer. Surface modification of the nanoparticles is carried out in accordance to the procedure outlined by Natarajan et al. [43]. The choice of polymer and surface modification determine nature of interactions between nanoparticle and polymer matrix. Our analysis [44] has shown that nanoparticle-polymer compatibility, quantified by ratio of work of adhesion, determines the likelihood of deagglomeration during extrusion. Incompatible systems

such as amino modified silica in PMMA matrix experienced less deagglomeration as compared to compatible systems.

Nanocomposites with 2wt% filler loading were prepared in a Thermo Haake Minilab, co-rotating twin screw extruder. Mixing parameters such as screw speed and specific energy input were varied to obtain a range of different dispersion states. A JEOL 2010 transmission electron microscope (TEM) was used to characterize the dispersion state of the nanocomposites. The TEM images were binarized using the Niblack algorithm [45, 46]. Dielectric spectroscopy measurements was carried out for each nanocomposites sample prepared for this study, details of which is available in ref. [44].

3.2 Microstructure Characterization and Reconstruction (Modules 1 & 3)

MCR enables extraction and quantitative representation of nanoparticle dispersion from TEM images of nanocomposites. The extracted representation will serve as microstructure parameters in PSP mapping and design optimization. In this article, dispersion is extracted using SDF, a frequency domain microstructure representation capable of capturing spatial correlations of complex heterogeneous materials. Mathematically, SDF $\rho(k)$ can be evaluated as:

$$\rho(\boldsymbol{k}) = |\mathcal{F}\{\mathcal{M}\}|^2, \tag{1}$$

where $\mathcal{M}$ is the binarized microstructure, $\mathcal{F}(.)$ is the Fourier transform operator and $\boldsymbol{k}$ is the frequency vector. For isotropic microstructures, SDF can be radially averaged about zero frequency such that the frequency vector $\boldsymbol{k}$ is reduced to a scalar k; making SDF a one-dimensional function of frequency. Although it is known to be the Fourier transform of a two-point autocorrelation function and hence encapsulates equivalent morphological information, Yu et al. [18] have shown that SDF is a more convenient representation to parametrize and design microstructures. These features are also evident from the analysis of nanocomposite microstructures in our database (Module 1). After binarizing TEM images using the Niblack algorithm [46] and assuming isotropy, SDF was evaluated using Eq.(1). We noticed that the SDF of all microstructures approximately follows an exponential distribution that can be parametrized with two variables – shape parameter α and scale parameter θ:

$$\rho(k) = \alpha * \exp\left(-\frac{k}{\theta}\right). \tag{2}$$

TEM images gathered from samples subjected to different processing conditions were characterized using SDF and parameters α and θ were ascertained by curve fitting using Eq.(2).

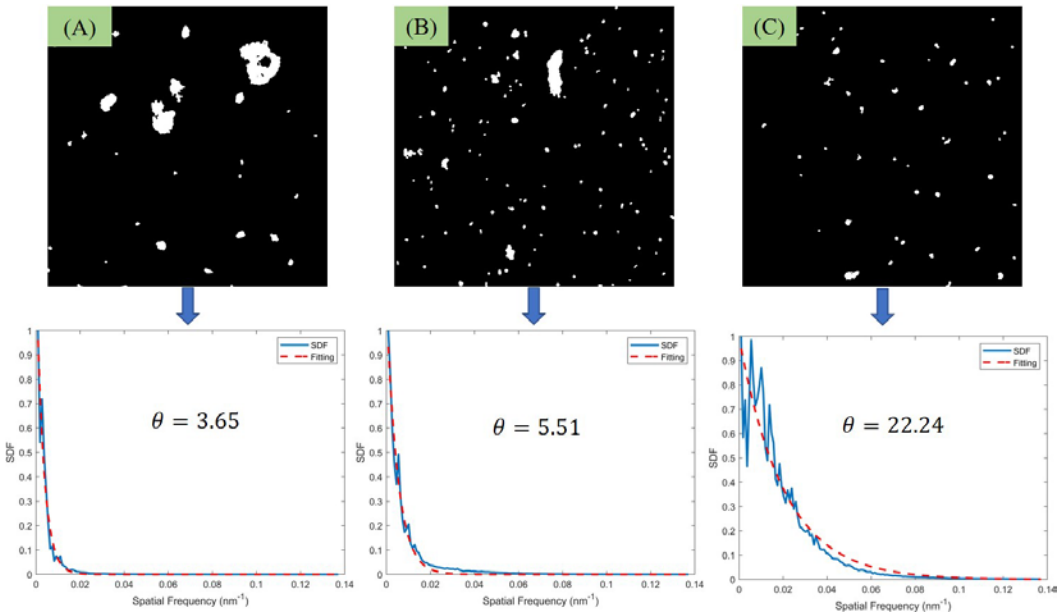

Figure 2: Three representative microstructures with varying dispersions and their SDF (blue curve) and corresponding curve fit using Eq. 2(red dashed curve). The design variable θ's value for each image shown in inset

The average R^2 value for fitting was 0.90. Images with exceptionally large nanoparticle agglomerates are not considered for this analysis as they do not significantly impact bulk nanocomposite response for loss or permittivity. Fig. 2 shows three microstructures along with their one-dimensional SDF and curve fitting. Filler dispersion increases through Fig. 2(A-C) and is reflected in a slower decay rate of SDF which can be quantified by θ. Each nanocomposite sample is represented by the average values of α and θ estimated from the analysis of TEM images. It was noticed that α varies in a narrow interval [0.39, 1.84] and has very little influence on the SDF profile. On the other hand, scale parameter θ varies between [1.49, 46.85], changing the rate of decay of SDF and consequently characterizing the dispersion of the filler aggregates. Thus, we will consider θ as a microstructure design variable and fix α to its mean value 1.1. The range of θ identified here will be used to define bounds for these variables in design formulation.

Microstructure reconstruction is an integral part of material design framework, since material properties must be evaluated for the microstructure represented by design variables at each iteration of optimization (Module 3). In this study, we are using the fast Fourier transform based reconstruction method developed by Iyer et al. [39].

3.3 Interphase Calibration and Finite Element Analysis for Dielectric Permittivity and Loss (Modules 2 and 3)

Each objective function evaluation (Module 3) is accomplished via finite element (FE) computation of the effective dielectric permittivity and loss of an RVE constructed using microstructure descriptor (dispersion) and composition (polymer type and surface modification type) recommended by BO. Incorporating interphase material properties into each FE simulation corresponding to the given combination of polymer type and surface modification type is a necessary intermediate step between constructing an RVE and computing its dielectric response [11]. Generally, we specify the permittivity and loss of the interphase in the form of five shifting factors that are applied to the polymer properties in the frequency domain to generate the complete frequency domain interphase properties [3, 47]. Frequency dependent dielectric properties, real ($\varepsilon'(\omega)$) and imaginary ($\varepsilon"(\omega)$) permittivity, of a polymer are expressed as superposition of independent Debye functions with different relaxation time (τ_i) and intensity ($\Delta\varepsilon_i$)

$$\varepsilon'(\omega) = \varepsilon_\infty + \sum_{i=1}^{n} \frac{\Delta\varepsilon_i}{1+(\omega\tau_i)^2}, \tag{3}$$

$$\varepsilon"(\omega) = \sum_{i=1}^{n} \frac{\Delta\varepsilon_i \omega\tau_i}{1+(\omega\tau_i)^2}, \tag{4}$$

Shift factors $C, M_\alpha, S_\alpha, M_\beta, S_\beta$ (α and β relaxation modelled separately) scale polymer relaxation time (τ_i) and intensity ($\Delta\varepsilon_i$) to generate interphase relaxation time ($S_\alpha\tau_i$, $S_\beta\tau_i$) and interphase intensity ($M_\alpha\Delta\varepsilon_i$, $M_\beta\Delta\varepsilon_i$). Superposition of Debye functions, as shown below, gives frequency dependent interphase properties

$$\varepsilon'_{int}(\omega) = \varepsilon_\infty + C + M_\alpha \sum_{\tau_i>\tau_0} \frac{\Delta\varepsilon_i}{1+(\omega S_\alpha\tau_i)^2} + M_\beta \sum_{\tau_i<\tau_0} \frac{\Delta\varepsilon_i}{1+(\omega S_\beta\tau_i)^2}, \tag{5}$$

$$\varepsilon"_{int}(\omega) = M_\alpha \sum_{\tau_i>\tau_0} \frac{\Delta\varepsilon_i \omega S_\alpha\tau_i}{1+(\omega S_\alpha\tau_i)^2} + M_\beta \sum_{\tau_i<\tau_0} \frac{\Delta\varepsilon_i \omega S_\beta\tau_i}{1+(\omega S_\beta\tau_i)^2}, \tag{6}$$

where τ_0, relaxation time corresponding to critical frequency, is used to make distinction between low frequency (α) and high frequency (β) regime. More details can be found in [47,48].

In this study, we focus on the design problem at a specific frequency target, 60Hz. Therefore, the calibration problem reduces from the task of finding five shifting factors to finding

two scale factors. These scale factors (SF_{real}, SF_{imag}) simply scale the polymer permittivity(ε') and loss (ε") at 60Hz to generate the corresponding interphase properties (${\varepsilon'}_{int}, {\varepsilon"}_{int}$) at 60Hz.

$$ {\varepsilon'}_{int}(\omega = 60Hz) = SF_{real} * \varepsilon'(\omega = 60hz), \quad (7) $$

$$ {\varepsilon"}_{int}(\omega = 60Hz) = SF_{imag} * \varepsilon"(\omega = 60hz), \quad (8) $$

Calibration of these scale factors (Module 2) is performed to minimize difference between the dielectric spectroscopy response of the FE simulation and that measured in experiments at 60Hz, for each of the six material combinations that span the design space. This calibration can be accomplished either with manual tuning by trial and error iterations [3, 47] or using black-box optimization methods, for instance, adaptive sampling using Bayesian approach [48], the former being used here. The trial and error calibration approach begins with simulation of the two phase microstructure (no interphase) to obtain the initial error with respect to the composite values at 60Hz. Based on this error, an initial assumption on the scaling factors for the interphase is made and used as input in a three phase model (with interphase) and the new output properties are predicted in FE. The values of the scale parameters are then varied iteratively until the error between the FE predicted properties for the three phase composite and the experimental data is less than the target acceptable error. A similar manual procedure can be followed, with some additional considerations, while tuning frequency dependent interphase description as explained in [47].

The calibration protocol (module 2) is performed once for each of the six possible material combinations. The RVE construction for the FE simulation is based on a microstructure constructed by averaging microstructure descriptors across all processing conditions (30 TEM images per processing condition) for that composition. Since a single interphase property is

Table 1: Dielectric properties (relative to vacuum permittivity of 8.85×10^{-12} F/m [3]) of interphase and pure polymer at 60Hz

Polymer – Surface Modification	Permittivity	Loss
PMMA	3.44	0.170
PS	2.02	0.001
PMMA-Chloro	3.10	0.120
PS-Chloro	6.00	0.010
PMMA-Nitro	2.70	0.050
PS- Nitro	4.80	0.023
PMMA-Octyl	4.20	0.250
PS-Octyl	5.70	0.035

expected for each material combination, we select the most representative experimental response (from data across multiple processing conditions) for tuning the scale factors. These assumptions, while necessarily containing approximations on material response, are sufficient to demonstrate the nanocomposite design process. Notably, this study does not attempt to calibrate the interphase separately for each processing condition although we acknowledge such calibration across processing conditions or a predictive model of interphase properties should be explored in the future, towards the physical validation of a predicted design and can possibly be done using data available in NanoMine (Module 1). Table 1 lists dielectric properties of pure polymers obtained in spectroscopy experiments and scaled interphase properties obtained by manual tuning for each material combination. These calibrated interphase properties are then used in the design process to assign appropriate interphase values for each design iteration according to material composition.

3.4 Machine Learning for Breakdown Strength Prediction (Module 4)

Dielectric breakdown of nanocomposites is a complex phenomenon and requires atomic scale simulations to decode the complex interactions occurring in the interphase. As current atomistic models are immature, we use a random forest [40] model trained on experimental data for rapid evaluation of U_d as a function of material design variables during optimization. Random forest technique was chosen due to its ability to handle mixed-variables, superior computational efficiency and minimal possibility of overfitting. Training data comprised U_d measurement

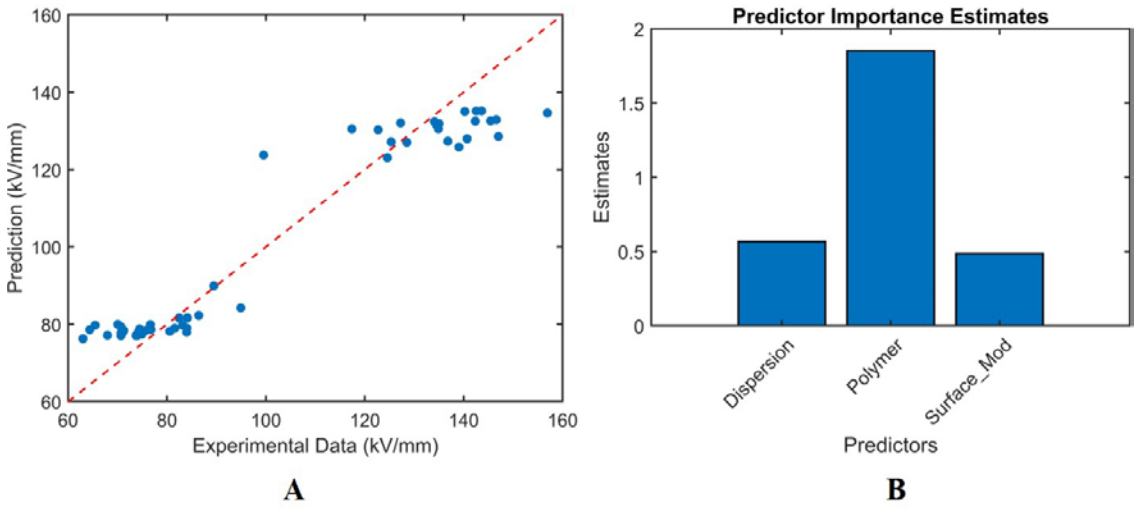

Figure 3: (A) Prediction accuracy of the random forest trained to predict breakdown strength. (B) Estimate of predictor importance deduced by random forest model. The larger the importance estimate for a predictor, the stronger its influence on breakdown strength.

(expressed in kV/mm) of 51 samples at 60 Hz. Predictors used for predicting U_d are the two qualitative (polymer type, surface modification type) and one quantitative (θ) design variables. A 10-fold cross validation study revealed that the random forest model with 500 trees predicts U_d accurately with a relative root mean square error of 0.38 and re-substitution $R^2 = 0.92$ (Fig. 3(A)). We observe the dataset to form two clusters; a PMMA based low U_d cluster and a PS based high U_d cluster. The strong influence of polymer is also confirmed by its large predictor importance estimate derived from the random forest model as shown in Fig. 3(B).

3.5 Latent Variable GP Modelling for Mixed-Variable Problems (Module 5)

One of the key components of BO is a statistical model that predicts the material properties from design variables and quantifies lack-of-data uncertainty. While Gaussian Processes (GP) are frequently used in BO, the standard GP methods were developed under the premise that all input variables are quantitative, which does not hold for concurrent composition and microstructure design of nanocomposite with two qualitative variables. We recently proposed using LVGP [26, 27] that maps the levels of the qualitative factor(s) to a set of numerical values for some latent quantitative variable(s). As illustrated in Fig. 4, our method is based on the belief that any qualitative factor must correspond to some underlying high-dimensional quantitative physical attributes that fully characterize that factor. Estimating the numerical latent variable values for the levels of the factor is essentially finding a mapping from the underlying high-dimensional space to the latent space, although we do not construct the mapping explicitly. The latent variables do

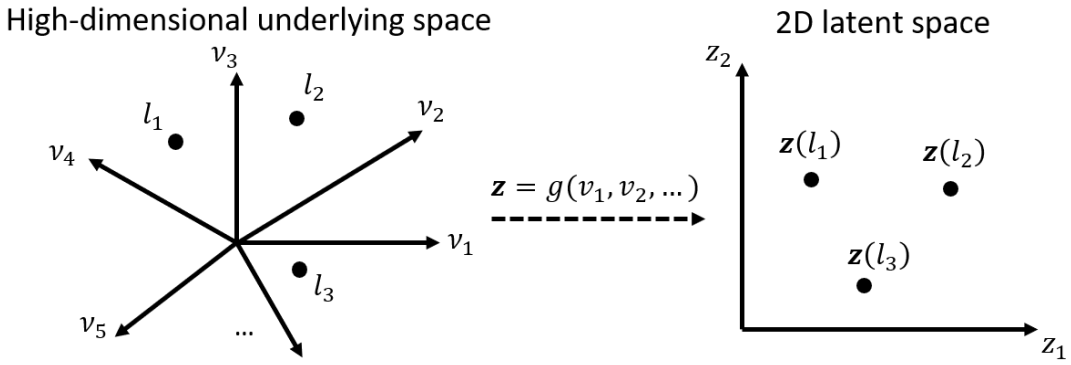

Figure 4: Illustration of high-dimensional underlying space of an arbitrary qualitative factor and the mapped latent space. The factor has levels l_1, l_2, and l_3, and is fully characterized by physical attributes v_1, v_2,…. The mapping $g: \boldsymbol{v} \rightarrow \boldsymbol{z}$ is implicitly constructed and found during the estimation of the latent variable values $\{\boldsymbol{z}(l_1), \boldsymbol{z}(l_2), \boldsymbol{z}(l_3)\}$.

not have explicit physical meanings, but they provide an inherent structure for the levels of the factor(s), which leads to substantial insight into the effects of the qualitative factors. For clarification, the latent variables are only used internally inside LVGP models. When LVGP models are used for predictions, they still take mixed-variable inputs in the original mixed-variable input spaces.

To describe the LVGP approach, the input variables are denoted as $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{t})$, where $\boldsymbol{x} = (x_1, x_2, \dots, x_p)$ represents p quantitative variables and $\boldsymbol{t} = (t_1, t_2, \dots, t_q)$ is the vector of q qualitative variables. With $i = 1,2, \dots, q$, the qualitative variable t_i has m_i levels $\{l_1^{(i)}, l_2^{(i)}, \dots, l_{m_i}^{(i)}\}$. The nanocomposite design problem under study has one quantitative variable (dispersion parameter θ), while the choice of polymer and surface modification are modeled as two qualitative variables with two (PMMA, PS) and three (Octyl, Chloro, Amino) levels respectively.

The output variable is denoted as y, and a set of data points of input-output pairs are noted as $\{(\boldsymbol{w}_1, y_1), \dots, (\boldsymbol{w}_N, y_N)\}$. In context of nanocomposites, the output variable can be one of the three dielectric properties. Then, consider the GP model

$$Y(\cdot) = \mu + G(\cdot), \tag{9}$$

where μ is the constant prior mean, and $G(\cdot)$ is a zero-mean GP with covariance function $k(\cdot,\cdot) = \sigma^2 r(\cdot,\cdot|\boldsymbol{\varphi})$. σ^2 is the prior variance of the GP, and $r(\cdot,\cdot|\boldsymbol{\varphi})$ is the correlation function parameterized with $\boldsymbol{\varphi}$. The true model $y(\cdot)$ is regarded as a realization of the GP $Y(\cdot)$. Once the form of the correlation function $r(\cdot,\cdot|\boldsymbol{\varphi})$ is specified, the hyperparameters $(\mu, \sigma^2, \boldsymbol{\varphi})$ can be estimated through maximum likelihood estimation (MLE) or other principles such as minimizing cross-validation errors. If the independent variables of the correction function $r(\cdot,\cdot|\boldsymbol{\varphi})$ are only the continuous variables $\boldsymbol{x}$, one can use the popular Gaussian correlation function

$$r(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\varphi}) = exp\left\{-\sum_{i=1}^{p} \varphi_i (x_i - x_i')^2\right\}, \tag{10}$$

which quantifies the correlation between $G(\boldsymbol{x})$ and $G(\boldsymbol{x}')$ for any input locations $\boldsymbol{x} = (x_1, \dots, x_p)$ and $\boldsymbol{x}' = (x_1', \dots, x_p')$ based on their 2-norm distance scaled by $\boldsymbol{\varphi}$. However, in the mixed-variable problem, it is not straightforward to incorporate the qualitative variable $\boldsymbol{t}$ in such a correlation function, as the difference $t_i - t_i'$ is undefined. The LVGP model handles this by mapping the qualitative variables $\boldsymbol{t}$ to quantitative ones.

In LVGP models, the m_i levels of the qualitative variable t_i are mapped to m_i latent numerical vectors $\{\boldsymbol{z}^{(i)}\left(l_1^{(i)}\right), \dots, \boldsymbol{z}^{(i)}\left(l_{m_i}^{(i)}\right)\}$ of a latent variable vector $\boldsymbol{z}^{(i)} \in \mathbb{R}^d$, where d is the dimensionality of $\boldsymbol{z}^{(i)}$. Modelers are free to choose the value of d as a modeling parameter, although setting $d = 2$ has been shown to be advisable for most problems. The original mixed-type input variables $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{t})$ are thus mapped to purely continuous variables $\left(\boldsymbol{x}, \boldsymbol{z}^{(1)}(t_1), \dots, \boldsymbol{z}^{(q)}(t_q)\right)$. A correlation function like Eq. (10) can be subsequently constructed as

$$r(\boldsymbol{w}, \boldsymbol{w}'|\boldsymbol{\varphi}, \boldsymbol{Z}) = exp\left\{-\sum_{i=1}^{p} \varphi_i\left(x_i - x_i'\right)^2 - \sum_{i=1}^{q} \left\|\boldsymbol{z}^{(i)}(t_i) - \boldsymbol{z}^{(i)}\left(t_i'\right)\right\|_2^2\right\}, \quad (11)$$

where $\boldsymbol{Z}$ is the collection of all the latent parameters $\{\boldsymbol{z}^{(1)}\left(l_1^{(1)}\right), \dots, \boldsymbol{z}^{(1)}\left(l_{m_1}^{(1)}\right), \boldsymbol{z}^{(2)}\left(l_1^{(2)}\right), \dots, \boldsymbol{z}^{(q)}\left(l_{m_q}^{(q)}\right)$. With this correlation structure, hyperparameters $(\mu, \sigma^2, \boldsymbol{\varphi}, \boldsymbol{Z})$ are obtained by MLE as in standard GP modelling. More details of this procedure and examples can be found in Zhang et.al [26].

LVGP serves as the machine learning model predicting the optimization objective(s) from the design variables i.e. Step I of the BO procedure described in Section 2. We use LVGP models with two-dimensional latent space representation for all optimization results reported in Section 4. Uncertainty quantification provided by LVGP is used accomplish Step II of the BO procedure as described below.

3.6 Bayesian Optimization (Module 5)

To meet the demand for electrical insulation, our goal is to identify nanocomposites with high U_d, low ϵ and low $tan\delta$. The design space consists of three variables, two qualitative and one quantitative, as summarized in Table 2. The choice of polymer and surface modification are qualitative variables with two (PS, PMMA) and three (Octyl, Chloro, Amino) levels respectively. Dispersion is a quantitative variable with bounds identified using SDF in Section 3.2. We present

Table 2: Summary of design variables used in case study

Variable	Type	Range/Levels
Polymer Type ($\boldsymbol{P}$)	Qualitative	$\{PMMA, PS\}$
Surface Modification Type ($\boldsymbol{S}$)	Qualitative	$\{Chloro, Octyl, Amino\}$
Filler Dispersion ($\boldsymbol{\theta}$)	Quantitative	[1.49,46.85]

both single and multicriteria BO strategies for this case study, using the same set of design variables with different objective formulations.

For single criterion BO, we formulate an objective function that weighs all three normalized properties (indicated by *) equally and adds/subtracts each property depending on whether it needs to be minimized (maximized):

$$\begin{gathered}\min_{s\in S,p\in P,m\in M} \tan\delta^* + \epsilon^* - \mathrm{U}_\mathrm{d}^* \\ S\colon \{Chloro, Octyl, Amino\} \\ P\colon \{PMMA, PS\} \\ M\colon microstructures\ with\ 1.49 \le \theta \le 46.85,\end{gathered} \tag{12}$$

where objective is to be minimized over a design space consisting of all possible combinations of surface modification (S), polymers (P) and microstructures (M). LVGP modeling is used to model the objective function with design variables S, P & M as inputs. Expected improvement [41] is used as the acquisition function due to its ability to balance exploration and exploitation of design space, thus converging to optimum rapidly. Eq. (12) can be modified by adding weights to each property expressing designer's priority for optimizing one property over the others. For example, maximizing U_d can be prioritized by assigning a weight factor of 10 in the objective function:

$$\min_{s\in S,p\in P,m\in M} \tan\delta^* + \epsilon^* - 10\mathrm{U}_\mathrm{d}^* \tag{13}$$

where S, P and M are the same as in Eq. (12). The modification of objective function subsequently affects the location of optimum in mixed-variable design space and will be discussed in Sec. 4.1.

Multicriteria optimization aims to find candidate designs lying on the Pareto frontier [49] – a characteristic boundary comprising designs where no criteria can be improved without the deterioration of others. The general multicriteria optimization problem can be formulated as

$$\min_{\boldsymbol{w}\in W}\{y_1(\boldsymbol{w}), y_2(\boldsymbol{w}), \ldots, y_s(\boldsymbol{w})\}, \tag{14}$$

where $\boldsymbol{w}$ is the design input, W is the design space, s is the number of criterion, and $\{y_1(\cdot), y_2(\cdot), \ldots, y_s(\cdot)\}$ is the set of the criterions that share the same design inputs. To identify the Pareto frontier for Eq. (14) numerically, the criteria are evaluated at a certain number of design inputs. Of all the evaluated design points, one selects the set of design points that are not dominated by any others. Here, a design point $\boldsymbol{w}$ is not dominated by another one $\boldsymbol{w}'$ if there exists at least one $i \in \{1,2,\ldots,s\}$ such that $y_i(\boldsymbol{w}) < y_i(\boldsymbol{w}')$. This set of design points is regarded as a representation of the true Pareto set.

To implement the BO approach for the multicriteria problem in Eq. (14), we use the expected maximin improvement (EMI) [42] acquisition function described as follows. Let the current Pareto set be composed of input set $P_W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_k\}$ and output set $P_Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_k\}$, where k is the number of points in the Pareto set and $\boldsymbol{y}_i = [y_1(\boldsymbol{w}_i), y_2(\boldsymbol{w}_i), \dots, y_s(\boldsymbol{w}_i)]^T, i = 1,2, \dots, k$. For any given new input $\boldsymbol{w}_0$, the corresponding outputs are predicted by the LVGP models as $\boldsymbol{Y}_0(\boldsymbol{w}_0) = [Y_1(\boldsymbol{w}_0), Y_2(\boldsymbol{w}_0), \dots, Y_s(\boldsymbol{w}_0)]^T$, where $Y_j(\boldsymbol{w}_0), j = 1,2, \dots, s$ is a random variable. To quantify how much the random outputs $\boldsymbol{Y}_0(\boldsymbol{w}_0)$ would improve the current Pareto set, we use the minimax improvement metric

$$I\big(\boldsymbol{Y}_0(\boldsymbol{w}_0)\big) = \min_{\boldsymbol{w}_i \in P_W} \left\{ max\left(\left\{ y_j(\boldsymbol{w}_i) - Y_j(\boldsymbol{w}_0) \right\}_{j=1}^{s} \cup \{0\} \right) \right\}, \tag{15}$$

which is also a random variable. The larger the value of $I\big(\boldsymbol{Y}_0(\boldsymbol{w}_0)\big)$ is, the more improvement the output $\boldsymbol{Y}_0(\boldsymbol{w}_0)$ is considered to make.

With this formula, if the output $\boldsymbol{Y}_0(\boldsymbol{w}_0)$ would be dominated by at least one point in the current Pareto set, then $I(\boldsymbol{Y}_0(\boldsymbol{w}_0)) = 0$, which means no improvement. Otherwise, $I(\boldsymbol{Y}_0(\boldsymbol{w}_0))$ would be a positive value quantifying the improvement. The value of $I(\boldsymbol{Y}_0(\boldsymbol{x}_0))$ is illustrated by a two-criteria example case in Fig. 5, with one of the candidate points being $I(\boldsymbol{Y}_0) = 0$ and the other two points with a positive value $I(\boldsymbol{Y}_0)$.

The criterion for choosing the new evaluation input $\boldsymbol{w}_0^*$ is to maximize the expected value of improvement given in Eq. (15), i.e.,

$$\boldsymbol{w}_0^* = \underset{\boldsymbol{w}_0 \in W}{argmax} E(I(\boldsymbol{Y}_0(\boldsymbol{w}_0))). \tag{16}$$

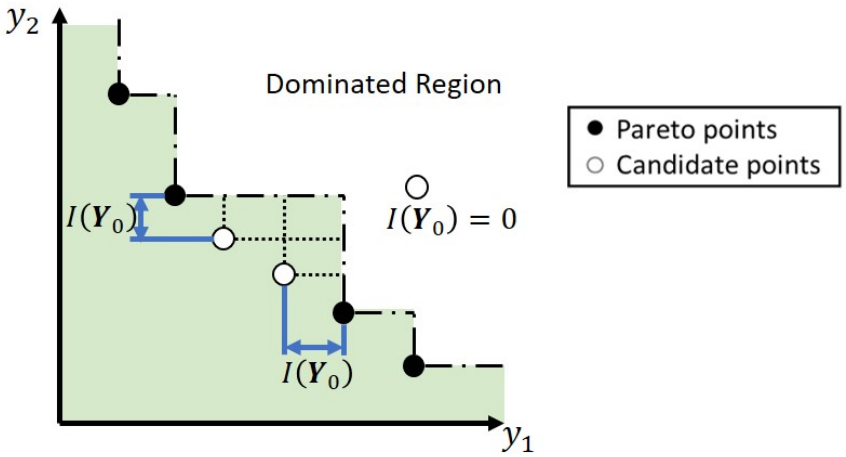

Figure 5: Values of the improvement metric $I(\boldsymbol{Y}_0)$ in a sampling process with two criteria.

When the original problem (Eq. (14)) has mixed-variable input space W, Eq. (16) is a mixed-variable optimization problem. To solve Eq. (16), we use a zero-order optimization strategy, where we generate a large set of candidate points in the input space, and then choose the one with the largest EMI as $\boldsymbol{w}_0^*$. For evaluating the expectation in Eq. (16), we use Monte Carlo simulation, as the analytical formula for EMI is too complex when $s \geq 3$, which is the case for nanocomposite design problem discussed here.

With three dielectric properties of interest, Eq. (14) is adapted for multicriteria nanocomposite design as follows:

$$\begin{gathered} \min_{\mathrm{s}\in\mathrm{S},\mathrm{p}\in\mathrm{P},\mathrm{m}\in\mathrm{M}} \tan\delta, \epsilon, -\mathrm{U_d}\ , \\ S\colon \{Chloro, Octyl, Amino\} \\ P\colon \{PMMA, PS\} \\ M\colon microstructures\ with\ 1.49 \leq \theta \leq 46.85, \end{gathered} \tag{17}$$

Where the variables have the same meaning as in Eq. (12). We use three independent LVGP models to predict the three dielectric properties from design variables S, P and M.

4. Optimization Results and Discussion

We performed 35 and 70 iterations of BO for single and multicriteria formulations respectively, as specified by Eq. (12) and Eq. (17) respectively. Each BO is initiated with 30 random initial samples where the values of quantitative variable $\{\theta\}$ are generated by Latin hypercube design and qualitative variables, polymer and surface modification type are sampled uniformly.

4.1 Results from single criterion Bayesian Optimization

We performed ten replicates of single criterion BO and each replicate is initiated with 30 random samples. We observed that all replicates consistently converge to optimal design with the objective value being -0.562, which corresponds to the design $\{\theta = 1.49, P = PS, S = Octyl\}$ with material properties $tan\delta = 0.0018$, $\epsilon = 2.211$ and $U_d = 127.67\frac{kV}{mm}$. Fig. 6(A) shows optimization history for one replicate and depicts evolution of design during optimization. We observe that octyl-modified Silica nanoparticles in PS with low dispersion is ideal to meet our requirements of high U_d, low $tan\delta$ and ϵ. These findings are consistent with our previous investigations that found $tan\delta$ and ϵ increase with dispersion. Not surprisingly, the choice of

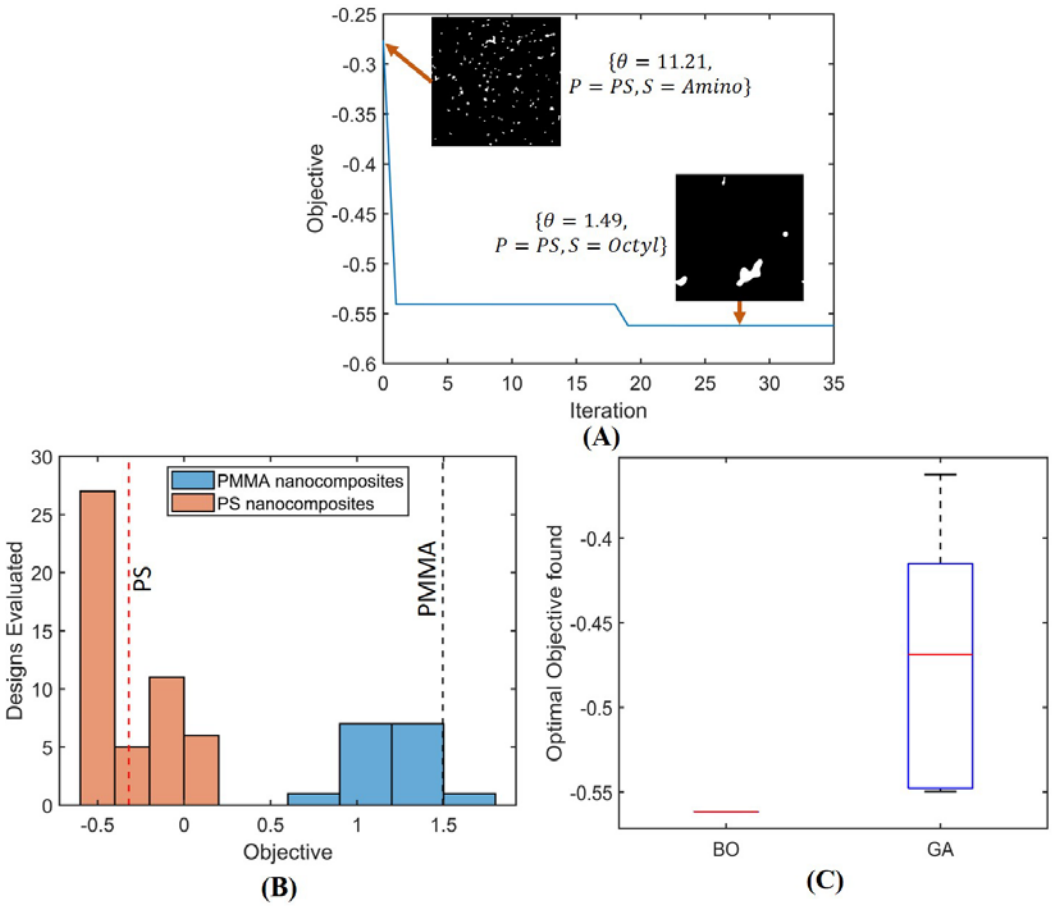

Figure 6: (A)Optimization history for single criterion BO that converged to objective = -0.562 along with three designs evaluated in the process (B) Distribution of evaluated designs, grouped by polymer type. Dashed lines denote objective values for PS & PMMA polymers (C) Comparison of ten replicates of BO and GA for single criterion optimization

polymer has a significant impact on the objective as indicated by Fig. 6(B). All PMMA based designs have large objective values compared to PS based designs. As a consequence, only 16 PMMA designs were evaluated in total (15 of which were provided in the dataset used for initialization) and BO strongly favored evaluation of PS based designs. We also notice that the objective value of optimum design (-0.562) shows a 75.9% improvement over pure PS properties (-0.319).

To demonstrate the efficacy of BO in identifying the optimal designs for problems with limited computational budget, we compare its performance against Genetic Algorithm (GA) [50]. MATLAB's implementation of GA for mixed integer optimization was used in this study and applied to problem formulation defined by Eq. (12). For a fair comparison with BO, GA was configured to terminate after 65 objective function evaluations (seven generations with a

population size of eight). Fig. 6(C) compares the optimal designs identified by 10 replicates of GA versus BO. We see that regardless of initial samples provided, BO can consistently converge to the optimum design while GA is highly susceptible to the initial population. This shows that the BO strategy of utilizing LVGP model uncertainty quantification to intelligently select new designs for evaluation makes it robust and faster at approaching global optimum compared with other algorithms that do not use this information.

We also performed optimization using Eq. (13) where U_d is assigned a weight factor of 10. In this case, BO converged to design $\{\theta = 13.52, P = PS, S = Amino\}$ with material properties $tan\delta = 0.0055,\ \epsilon = 2.888$ and $U_d = 134.601 \frac{kV}{mm}$. In comparison to optimal design found using Eq. (12), this design has higher U_d at the expense of higher $tan\delta$ and ϵ due to more disperse nanoparticles. This exercise demonstrates that approaching a multicriteria design problem using a single criterion optimization technique is sensitive to formulation of objective function.

4.2 Results from Multicriteria Bayesian Optimization (MBO)

70 iterations of MBO were performed starting with 30 random initial samples. Three independent LVGP models are used to evaluate the three criteria. Fig. 7 displays the 2D latent space for two categorical variables – choices of polymer and surface modification for the LVGP models used in multicriteria optimization. LVGP constrains the first category (PMMA for polymers, Octyl for surface modification) to the origin and second category (PS for polymer, Chloro for surface modification) to the z_1 axis. The Euclidean distance between categories is used to calculate the correlation function as indicated in Eq. (11).

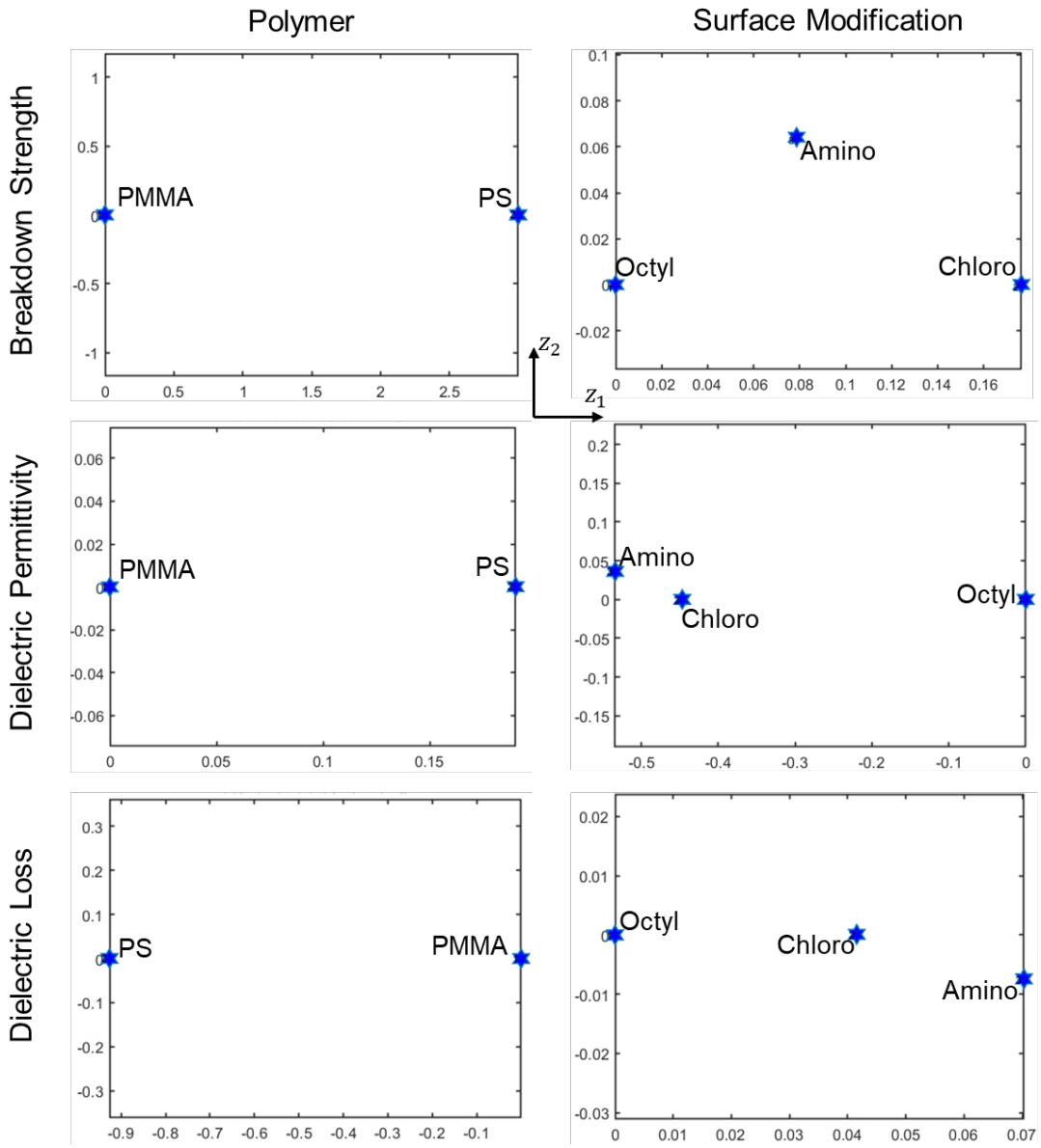

Figure 7: Visualization of latent variables for polymer and surface modification variables. Each row represents the latent variables estimated by the LVGP model used for corresponding property.

Fig. 8 plots the random initial samples and 16 designs that were identified on the Pareto front. A noticeable feature in this plot is that the initial samples create two clusters corresponding to two polymers under consideration. The cluster located in the low U_d, high $tan\delta$ and ϵ region (top left corner in Fig.8) exclusively contains PMMA based samples and is not favorable to meet the design criteria. This is consistent with the findings in Fig. 6(B). On the other hand, PS-based samples have higher U_d, lower $tan\delta$ and ϵ; suggesting that they are better suited for electrical insulation application compared to PMMA samples. This is also reflected in the fact that designs

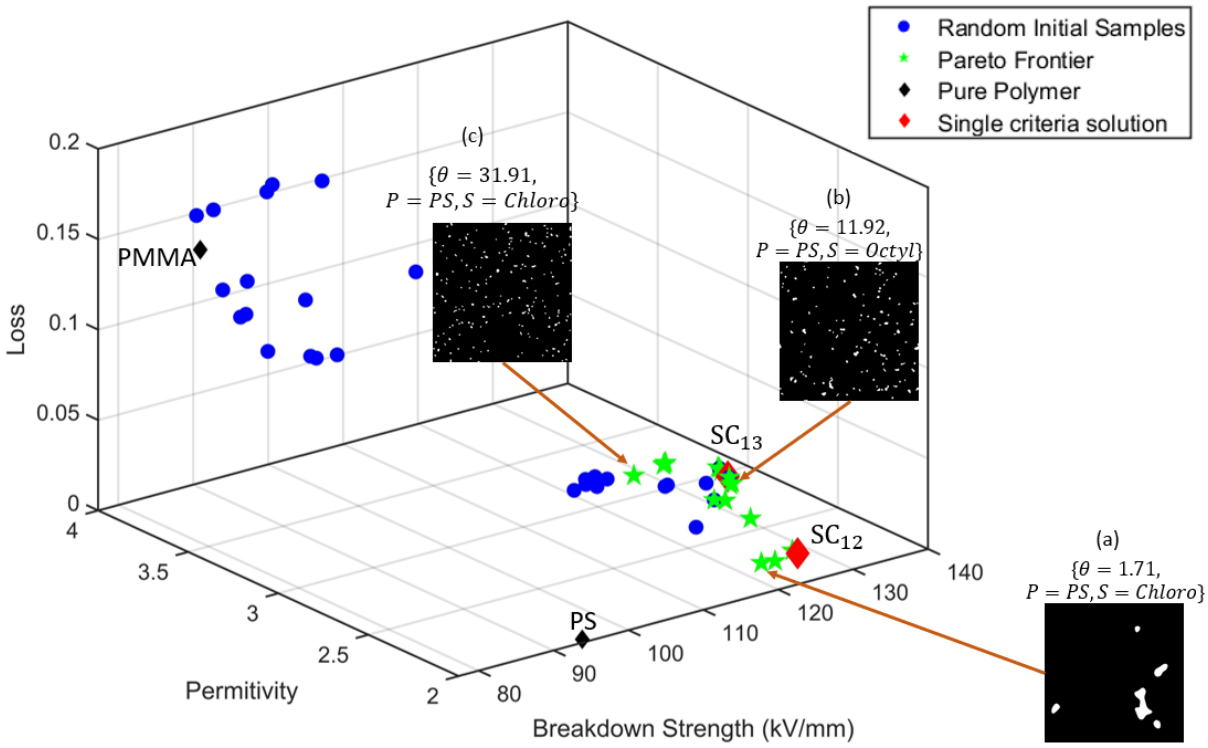

Figure 8: Summary of 70 iterations of Multicriteria Bayesian Optimization. SC_{12} and SC_{13} denote optimal single criterion solutions identified from Eq. (12) and Eq. (13) respectively.

evaluated by MBO are predominantly PS based. Notice that the Pareto front obtained by MBO shows significant improvement with regard to random initial samples and thus underlines the capability of uncertainty driven MBO to locate improved designs. The two optimal designs identified by single criterion BO are located in different regions of the Pareto front. While we had to repeat single criterion BO with different objective formulations, one simulation of MBO discovers these designs automatically to present the modeler with a diverse set of designs for consideration.

The influence of design variables on dielectric properties via Fig. 9, which displays the properties of 16 Pareto front identified by MBO. Compared to pure PS properties, PS based nanocomposites have higher dielectric properties values. These properties are also positively correlated to θ; they increase as dispersion increases. However, the rate of increases decreases beyond $\theta \sim 15$. While Chloro modification is ideal for minimizing $tan\delta$, it also contributes to higher ϵ. On the other hand, designs with Octyl and Amino surface modifications have lower ϵ but higher $tan\delta$ as compared to those with Chloro surface modification. Thus, we see a tradeoff between the three properties of interest. Selecting one among the several Pareto front designs for

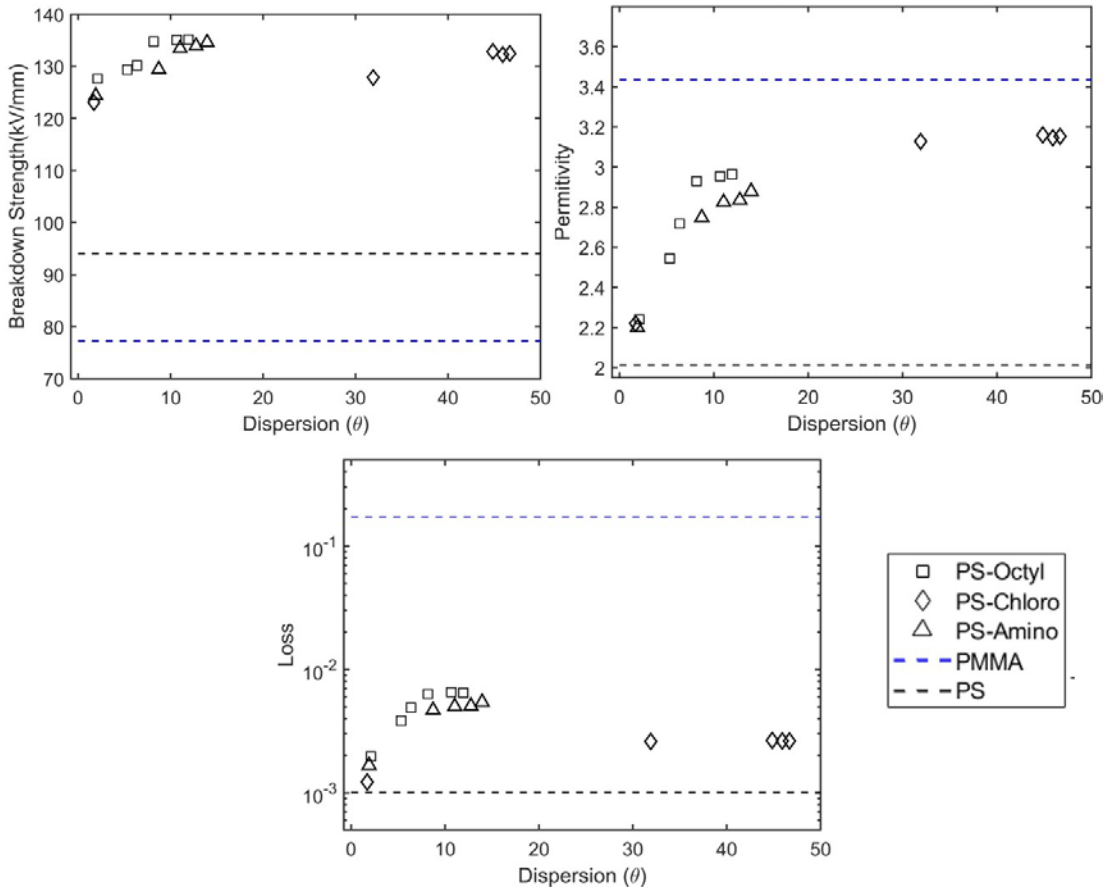

Figure 9: Influence of design variables on dielectric properties of nanocomposites on Pareto front. Dashed lines indicate property of polymer only system.

detailed analysis and testing depends on the modeler's preference based on the application, how the material is deployed, and device level performance.

Once the optimal design is identified, the corresponding processing condition can be obtained by mapping the optimized design variables to processing energy using the PS relationship established in our previous work [45]:

$$\bar{\mathrm{I}}_{\mathrm{filler}} = \mathrm{f(matrix)}\,\sinh^2(2\,\mathrm{W_{PF}}/\mathrm{W_{FF}} - 1)\log(\mathrm{E_\gamma} + 1) + C_0, \qquad (18)$$

where $\bar{\mathrm{I}}_{\mathrm{filler}}$ is the normalized interphase area, $\mathrm{f(matrix)}$ and $\mathrm{C_0}$ are polymer dependent constants, $\mathrm{W_{PF}}/\mathrm{W_{FF}}$ is the filler-matrix compatibility descriptors and $\mathrm{E_\gamma}$ is the processing energy descriptor that we seek. For illustration, we choose the design (b) in Fig. 8, favoring high breakdown strength, as our optimal solution. Microstructure reconstruction corresponding to $\theta = 11.92$ was performed and $\bar{I}_{filler}$ was found to be 0.189. For PS, $f(\mathrm{matrix})$ and $\mathrm{C_0}$ are 0.00995 and 0.08798 respectively. For octyl-modified silica nanoparticles dispersed in PS, $\mathrm{W_{PF}}/\mathrm{W_{FF}} = 1.15$. Plugging these values

in Eq. (18) leads to $E_\gamma = 32.77\ J/g$. Thus, we can identify designs satisfying application specific material properties and deduce processing parameter necessary for manufacturing.

5. CONCLUSIONS

This article presented a data-centric mixed-variable Bayesian Optimization framework for design of polymer nanocomposite with both qualitative and quantitative variables. Initiated by a nanocomposite database, our framework integrated empirical data with state-of-the-art techniques in interphase calibration, SDF based MCR for dimensionality reduction, and FEA-based structure-property simulations. Experimental property measurements are also leveraged for training machine learning models to predict material properties when theory based simulation models are lacking. Going beyond traditional BO implementations for quantitative design variables, mixed-variable modelling enabled by LVGP models allowed us to parsimoniously incorporate qualitative variables in a BO based design process. This capability is critical to accomplish concurrent composition and microstructure design which is inherently a mixed-variable optimization problem. Since functional materials must often meet multiple performance criteria, we extended LVGP based BO to multicriteria optimization using the expected maximin improvement acquisition function.

The efficacy of our data-centric framework was demonstrated through a case study focused on insulating nanocomposite design. The design formulation for single and multicriteria BO was presented using two qualitative (types of polymer and surface modification) and one quantitative (filler dispersion) variables. Modifying the weight assigned to breakdown strength demonstrated that single criterion BO is sensitive to objective formulation and does not have a unique solution when applied to multicriteria problems. On the other hand, multicriteria BO provides a variety of designs representing tradeoffs among dielectric properties, allowing the modeler to select a solution based on their preference. Processing energy required for fabrication of optimal design was evaluated using processing to structure mapping, to complete the bi-directional traversal across PSP paradigms and demonstrate the material genome approach to material design. While LVGP based BO is applicable to any engineering design problem, the unique ability to facilitate concurrent optimization of composition and microstructure w.r.t. one or more properties, makes it a powerful tool for materials design.

In the future, developing accurate simulation models based on Molecular Dynamics and Density Functional Theory is necessary for understanding and evaluating material properties such as dielectric breakdown strength and interphase behavior. Additionally, we are continuously

expanding NanoMine, the polymer nanocomposite data repository, by introducing standardized data curation workflows, data visualization capability and sophisticated interphase calibration and FEA tools described in this article. Several MCR methods including SDF are currently available in NanoMine[‡]. We envision NanoMine to drive the widespread adoption of data centric design methodology in the nanocomposite community.

ACKNOWLEDGEMENT

Support from NSF grants (ACI 1640840, CMMI 1729452, CMMI 1818574, CMMI 1729743, CMMI 1537641, OAC 1835782) and Center for Hierarchical Materials Design (ChiMaD NIST 70NANB14H012) are greatly appreciated.

CONFLICT OF INTEREST

There are no conflicts to declare.

REFERENCES

[1] J. P. Holdren, "Materials genome initiative for global competitiveness," *National Science and technology council OSTP. Washington, USA*, 2011.

[2] G. B. Olson, "Computational design of hierarchically structured materials," *Science*, vol. 277, no. 5330, pp. 1237-1242, 1997.

[3] H. Zhao *et al.*, "Dielectric spectroscopy analysis using viscoelasticity-inspired relaxation theory with finite element modeling," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 24, no. 6, pp. 3776-3785, 2017.

[4] Y. Huang *et al.*, "Predicting the breakdown strength and lifetime of nanocomposites using a multi-scale modeling approach," *Journal of Applied Physics*, vol. 122, no. 6, p. 065101, 2017.

[5] X. Li *et al.*, "Rethinking Interphase Representations for Modeling Viscoelastic Properties for Polymer Nanocomposites," *arXiv preprint arXiv:1811.06238*, 2018.

[6] J. S. Jang, B. Bouveret, J. Suhr, and R. F. Gibson, "Combined numerical/experimental investigation of particle diameter and interphase effects on coefficient of thermal expansion and young's modulus of SiO2/epoxy nanocomposites," *Polymer Composites*, vol. 33, no. 8, pp. 1415-1423, 2012.

[7] X. Cheng, K. W. Putz, C. D. Wood, and L. C. Brinson, "Characterization of local elastic modulus in confined polymer films via AFM indentation," *Macromolecular rapid communications*, vol. 36, no. 4, pp. 391-397, 2015.

[8] P. F. Brune *et al.*, "Direct Measurement of Rubber Interphase Stiffness," *Macromolecules*, vol. 49, no. 13, pp. 4909-4922, 2016.

[9] M. G. Todd and F. G. Shi, "Validation of a novel dielectric constant simulation model and the determination of its physical parameters," *Microelectronics journal*, vol. 33, no. 8, pp. 627-632, 2002.

[10] P. Maity, N. Gupta, V. Parameswaran, and S. Basu, "On the size and dielectric properties of the interphase in epoxy-alumina nanocomposite," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 17, no. 6, 2010.

[‡] NanoMine is accessible at https://materialsmine.org/nm#/

[11] R. Qiao and L. C. Brinson, "Simulation of interphase percolation and gradients in polymer nanocomposites," *Composites Science and Technology,* vol. 69, no. 3, pp. 491-499, 2009.

[12] R. Bostanabad *et al.*, "Computational Microstructure Characterization and Reconstruction: Review of the State-of-the-art Techniques," *Progress in Materials Science,* 2018.

[13] H. Xu, D. A. Dikin, C. Burkhart, and W. Chen, "Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials," *Computational Materials Science,* vol. 85, pp. 206-216, 2014.

[14] H. Xu, Y. Li, C. Brinson, and W. Chen, "A descriptor-based design methodology for developing heterogeneous microstructural materials system," *Journal of Mechanical Design,* vol. 136, no. 5, p. 051007, 2014.

[15] U. Farooq Ghumman *et al.*, "A Spectral Density Function Approach for Active Layer Design of Organic Photovoltaic Cells," *Journal of Mechanical Design,* vol. 140, no. 11, pp. 111408-111408-14, 2018.

[16] S. Torquato, *Random heterogeneous materials: microstructure and macroscopic properties*. Springer Science & Business Media, 2013.

[17] D. Chen and S. Torquato, "Designing disordered hyperuniform two-phase materials with novel physical properties," *Acta Materialia,* vol. 142, pp. 152-161, 2018.

[18] S. Yu *et al.*, "Characterization and design of functional quasi-random nanostructured materials using spectral density function," *Journal of Mechanical Design,* vol. 139, no. 7, p. 071401, 2017.

[19] H. Zhao, X. Li, Y. Zhang, L. S. Schadler, W. Chen, and L. C. Brinson, "Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design," *APL Materials,* vol. 4, no. 5, p. 053204, 2016.

[20] H. Zhao *et al.*, "NanoMine schema: An extensible data representation for polymer nanocomposites," *APL Materials,* vol. 6, no. 11, p. 111108, 2018.

[21] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE,* vol. 104, no. 1, pp. 148-175, 2016.

[22] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization,* vol. 13, no. 4, pp. 455-492, 1998.

[23] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, and T. Lookman, "Adaptive strategies for materials design using uncertainties," *Scientific reports,* vol. 6, p. 19660, 2016.

[24] C. Li *et al.*, "Rapid Bayesian optimisation for synthesis of short polymer fiber materials," *Scientific Reports,* vol. 7, no. 1, p. 5683, 2017/07/18 2017.

[25] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, "Crystal structure prediction accelerated by Bayesian optimization," *Physical Review Materials,* vol. 2, no. 1, p. 013803, 2018.

[26] Y. Zhang, S. Tao, W. Chen, and D. W. Apley, "A latent variable approach to Gaussian process modeling with qualitative and quantitative factors," *Technometrics,* pp. 1-12, 2019.

[27] Y. Zhang, D. W. Apley, and W. Chen, "Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables," *Scientific Reports,* vol. 10, no. 1, pp. 1-13, 2020.

[28] P. Akcora *et al.*, "Anisotropic self-assembly of spherical polymer-grafted nanoparticles," *Nature materials,* vol. 8, no. 4, p. 354, 2009.

[29] S. K. Kumar, N. Jouault, B. Benicewicz, and T. Neely, "Nanocomposites with polymer grafted nanoparticles," *Macromolecules,* vol. 46, no. 9, pp. 3199-3214, 2013.

[30] G. Munaò *et al.*, "Molecular structure and multi-body potential of mean force in silica-polystyrene nanocomposites," *Nanoscale,* vol. 10, no. 46, pp. 21656-21670, 2018.

[31] G. Munaò, A. De Nicola, F. Müller-Plathe, T. Kawakatsu, A. Kalogirou, and G. Milano, "Influence of Polymer Bidispersity on the Effective Particle–Particle Interactions in Polymer Nanocomposites," *Macromolecules,* vol. 52, no. 22, pp. 8826-8839, 2019.

[32] V. Ganesan and A. Jayaraman, "Theory and simulation studies of effective interactions, phase behavior and morphology in polymer nanocomposites," *Soft Matter,* vol. 10, no. 1, pp. 13-38, 2014.

[33] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science," *Apl Materials,* vol. 4, no. 5, p. 053208, 2016.

[34] Y. Zhang, H. Zhao, I. Hassinger, L. Brinson, L. Schadler, and W. Chen, "Microstructure reconstruction and structural equation modeling for computational design of nanodielectrics," *Integrating Materials and Manufacturing Innovation,* vol. 4, no. 1, p. 14, 2015.

[35] W. Chen *et al.*, "Materials Informatics and Data System for Polymer Nanocomposites Analysis and Design," in *Handbook on Big Data and Machine Learning in the Physical Sciences*, pp. 65-125.

[36] J. R. Weidner, F. Pohlmann, P. Gröppel, and T. Hildinger, "Nanotechnology in high voltage insulation systems for turbine generators-First results," *17th ISH, Hannover, Germany,* 2011.

[37] J. W. McPherson, J. Kim, A. Shanware, H. Mogul, and J. Rodriguez, "Trends in the ultimate breakdown strength of high dielectric-constant materials," *IEEE transactions on electron devices,* vol. 50, no. 8, pp. 1771-1778, 2003.

[38] Wei Chen *et al.*, "Materials Informatics and Data System for Polymer Nanocomposites Analysis and Design," in *Big, Deep, and Smart Data in the Physical Sciences*, 2018.

[39] A. Iyer *et al.*, "Designing anisotropic microstructures with spectral density function," *Computational Materials Science*, vol. 179, p. 109559, 2020.

[40] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[41] J. Mockus, V. Tiesis, and A. Zilinskas, "The application of Bayesian methods for seeking the extremum," *Towards global optimization,* vol. 2, no. 117-129, p. 2, 1978.

[42] D. C. T. Bautista, "A sequential design for approximating the pareto front using the expected pareto improvement function," The Ohio State University, 2009.

[43] B. Natarajan, Y. Li, H. Deng, L. C. Brinson, and L. S. Schadler, "Effect of Interfacial Energetics on Dispersion and Glass Transition Temperature in Polymer Nanocomposites," *Macromolecules*, vol. 46, no. 7, pp. 2833-2841, Apr 2013.

[44] A. Prasad, "Processing-Structure-Property Relationship for Polymer Nanodielectrics," 2019.

[45] I. Hassinger *et al.*, "Toward the development of a quantitative tool for predicting dispersion of nanocomposites under non-equilibrium processing conditions," *Journal of Materials Science,* vol. 51, no. 9, pp. 4238-4249, May 2016.

[46] W. Niblack, *An Introduction to Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986, pp. 115-116.

[47] Y. Huang *et al.*, "Prediction of interface dielectric relaxations in bimodal brush functionalized epoxy nanodielectrics by finite element analysis method," in *2014 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP)*, 2014, pp. 748-751: IEEE.

[48] Y. Wang *et al.*, "Identifying interphase properties in polymer nanocomposites using adaptive optimization," *Composites Science and Technology,* vol. 162, pp. 146-155, 2018.

[49] Y. Censor, "Pareto optimality in multiobjective problems," *Applied Mathematics and Optimization,* journal article vol. 4, no. 1, pp. 41-59, March 01 1977.

[50] D. E. Goldberg, *Genetic algorithms*. Pearson Education India, 2006.