NORTHWESTERN UNIVERSITY

Data Centric Design for Microstructural Materials Systems

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Mechanical Engineering

By

Akshay Iyer

EVANSTON, ILLINOIS

December 2021

ABSTRACT

Data Centric Design for Microstructural Materials Systems

Akshay Iyer

Materials science has been central to human advancement since time immemorial. There has always been curiosity around studying the processes required to extract materials, examine their structure, and ultimately tailor their properties to meet human needs. Over the last few centuries, the ability to tailor material properties was driven by design rules identified via experimentation, theoretical analysis, and more recently computational capabilities. It is only over the last decade that we have realized the immense potential of data driven materials discovery. This dissertation further examines this new paradigm of material discovery through the lens of design engineering. We show that the decision made during material design process – design representation, design evaluation and design synthesis are informed by the process-structure-property knowledge contained in material databases. Through a variety of advanced material systems, we seek to address some challenges arising at the intersection of design engineering and material science.

We investigate the design representation challenges arising in microstructure design. Spectral Density Function (SDF), a frequency domain microstructure approach is the focus of our study. We present a computational microstructure design framework for Organic Photovoltaic Cells (OPVC) using SDF and a novel structure-property simulation model. After identifying that there is a lack of microstructure representation and design methodologies for anisotropic microstructures, we demonstrate that SDF is capable of capturing the necessary information. Since

design requires reconstruction of microstructures, we present a novel methodology to reconstruct isotropic and anisotropic microstructures. Our method is also computationally efficient than the existing ones. Finally, we show that this capability is useful for designing the active layer of OPVCs that outperform their isotropic contemporaries.

The ability to design microstructures is useful for a wide variety of material systems, including polymer nanocomposites. In addition to the microstructure, the choice of constituents (polymer, filler and the filler's surface modification) have a significant influence on behavior of nanocomposites. Consequently, we cast the nanocomposite design as a mixed variable, multicriteria optimization problem and leverage Latent Variable Gaussian Processes (LVGP) and Bayesian Optimization (BO) to identify Pareto optimal candidates for electric insulation. This design methodology involves usage of experimental datasets for calibrating physics models, training property prediction models as well identifying the bounds for design variables.

The material properties are seldom determined completely by the composition. One such example is the metal insulator transition (MIT) compounds which display abrupt changes in their resistivity. To make them viable as next generation microelectronic devices, there is a growing interest in identifying compositions that simultaneously induce a large bandgap and high stability. We show that this combinatorial multicriteria optimization can be solved efficiently using LVGP and BO. LVGP allows us the circumvent the conventional feature engineering stage of design process which is extremely challenging for MIT due to limited understanding of the underlying physics.

Although qualitative variables encountered in nanocomposite and MIT design have few levels, some material systems may involve high dimensional qualitative variables i.e., with many levels.

This scenario poses a significant increase in computational cost of initiating BO since each level of every qualitative variable must be observed at least once for its latent variables to be estimated by LVGP. To this end, we develop a descriptor aided BO methodology that allows us to initiate BO with a small dataset (~$O(1)$) and parsimoniously predict latent variables for unobserved levels. The method is inspired by the belief that effect of qualitative variables is described by underlying numerical descriptors. Through a variety of examples, we outline the efficacy of our method in tackling several scenarios of partial and imperfect descriptor knowledge encountered in real world applications.

While the critical role of microstructure in material design is acknowledged in the research community, the computational Microstructure Characterization and Reconstruction techniques are not easily accessible. To this end, we have developed eight webtools with friendly graphical user interface in NanoMine to allow users to analyze their microstructural images with only a few clicks of the button.

Through a variety of material systems, this thesis exemplifies the strong confluence of material science with design engineering as outlined in the data centric design framework. With ever increasing focus on large scale data collection and analysis, we believe this framework serves as a guide to researchers for identifying critical tasks vis-à-vis data collection, method selection and method development required in the materials design process.

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude towards my PhD advisor Prof. Wei Chen. Throughout the many ups and downs of my PhD journey, she has been extremely supportive of my research endeavors and professional development activities. Her insights into research tasks, ability to work patiently with students and project management skills are exemplary and have taught me valuable skills to take beyond graduate school. I would also like to thank my dissertation committee members Prof. James Rondinelli, Prof. Daniel Apley and Prof. Cheng Sun for their helpful suggestions to the various chapters of this thesis. Prof. Rondinelli's wordsmanship and figure making skill were pivotal in the publication of the article on electronic materials discovery in Applied Physics Review. Dr. Apley's timely critiques of my work have helped me analyze methods rigorously and refine them as appropriate. Despite the various challenges brought by COVID-19 induced Zoom lecturing, Prof. Sun's supportive nature made my life easier as a Teaching Assistant for his course Engineering Analysis 3.

Throughout my PhD journey, I have been fortunate to work with excellent collaborators from various domains. Their willingness to share datasets and provide insights into material design problems has helped me identify new research directions, some of which have been included in this thesis. Here, I express appreciation to these collaborators:

- Organic Photovoltaics: Prof. TeYu Chien, Rabindra Dulal, Prof. Ganesh Balasubramanian, Dr. Joydeep Munshi
- Polymer Nanocomposites: Prof. L.Catherine Brinson, Prof. Linda Schadler, Prof. Ravishankar Sundararaman, Praveen Gupta, Abhishek Shandilya, Prajakta Prabhune, Dr. Boran Ma, Dr. Yixing Wang, Dr. Aditya Prasad

- Metal Insulator Transition materials: Dr. Raymond Wang, Dr. Alexandru Georgescu, Suraj Yerramilli

The IDEAL laboratory has provided me a helpful and collaborative ecosystem to thrive in. Special thanks to my colleagues Dr. Siyu Tao, Anton van Beek and Yu-Chin Chan who are always willing to have lengthy conversations about research, coursework, job hunt and the various vicissitudes of life (although these conversations would often just involve me complaining about things). Dr. Yichi Zhang, Dr. Xiaolin Li and Dr. Ramin Bostanabad were extremely supportive during my formative years in the lab. Their suggestions and advice on a variety of issues have been greatly beneficial. In addition, my interactions with Dr. Tianyu Huang, Umar Ghumman, Liwei Wang, Henry Zhang, Yigitcan Comlek, Yaxin Cui, Mouad Fergoug, Dr. Daicong Da and Dr. Faez Ahmed have been enjoyable.

Being part of the Mechanical Engineering Graduate Student Society (MEGSS) was a joyful experience and provided me a chance to make friends outside my lab. I would be remiss if I did not thank fellow MEGSS members Suman Bhandari, Lichao Fang, Aleksandra Kalinowska, Hannah Emnett and Thomas Janssen for helping me organize various events over my two year stint with MEGSS.

Finally, I would like to thank my family for their unconditional support throughout my education.

To my parents, who sacrificed many comforts of life to prioritize their children's education.

Contents

List of Figures

List of Tables

1 Research Motivation and Objectives

1.1 Data Centric Framework for Material Design

Design Engineering can be described as an amalgamation of Design Representation, Design Evaluation and Design Synthesis. Design Representation encompasses methods that characterize the control factors i.e., the variables that influence system behavior. Design Evaluation entails methodology to evaluate system response from its representation. Both are heavily dependent on the system being studied. The knowledge gained from these tasks is utilized in the final step of design process - Design Synthesis. It involves navigating the design space to identify optimal designs. While significant developments have taken place in design engineering over the last century, recent interest in application of these methods for advanced materials development reveals new challenges.

For most of 20th century, material science research and development were performed through Edisonian "trial and error" approach, which is time consuming, expensive, and often delayed the deployment of emerging materials in commercial applications. To bring about changes in the way we design materials, there is a need to shift the focus of material science research from simply being an explanation of observed phenomena to development of predictive models that identify the underlying factors controlling the phenomena and tuning them to meet the desired objectives for industrial applications. This has been the theme of Material Genome Initiative (MGI) [1], which has revolutionized the way advanced material systems are designed with targeted performance. MGI strives at elucidating the relationship between Processing-Structure-Property (PSP) [2] paradigms for material design. It requires development of new methods within each of the three

domains and protocols to manage information flow across domains. A holistic design strategy for bi-directional traversal of PSP relations requires us to address certain key issues – cost effective processing techniques, microstructure representation and reconstruction, dimensionality reduction and tractable optimization techniques. Recently, the emergence of open-source material databases [3-7] and gaining popularity of machine learning techniques is accelerating our ability to address some of these challenges using a data-centric approach. NanoMine [3, 4], a nanocomposite material database with in-built data curation, exploration and analysis capabilities, represents this approach in the field of polymer nanocomposites. It captures the physical properties reported in the literature and from individual research labs including microstructure, processing conditions, and material properties. Ontology-enabled knowledge graph framework helps NanoMine establish relationship between those properties. A collection of module tools for microstructure characterization & reconstruction and simulation software to model bulk nanocomposite material response augments knowledge generated by experimental data. Integrating these different sources of knowledge is critical for material design. However, generating experimental or simulated data for the vast design space defined by the almost infinite combinations of constituents, microstructure morphology, and processing conditions is impractical. This signifies the need for data-centric methodologies that can effectively interrogate existing data and interpolate between them to find new high performing materials.

To this end, we present a data centric framework for material design (Figure 1-1) where each step of the design process is guided by knowledge stored in databases. The choice of design representation is dependent on domain knowledge about factors known to influence material property. For example, bandgap of inorganic compounds is entirely determined by its composition

and thus composition is itself a suitable representation. On the other, electrical properties of polymer nanocomposites depends on composition as well as microstructure and necessitates the use of low dimensional microstructure representation methods such as Correlation Functions, Spectral Density Function (SDF), Physical Descriptors etc. Evaluating material property from its representation is heavily reliant on length & time scales at which the underlying phenomena takes place. For instance, Density Functional Theory calculations capture atomic level properties such as band gap, Molecular Dynamics simulations model an ensemble of molecules while finite element analysis is suitable for phenomena occurring at higher length scales. Each of these methods require calibration of embedded parameters & validation of property predictions, which is accomplished through experimental data contained in the database.

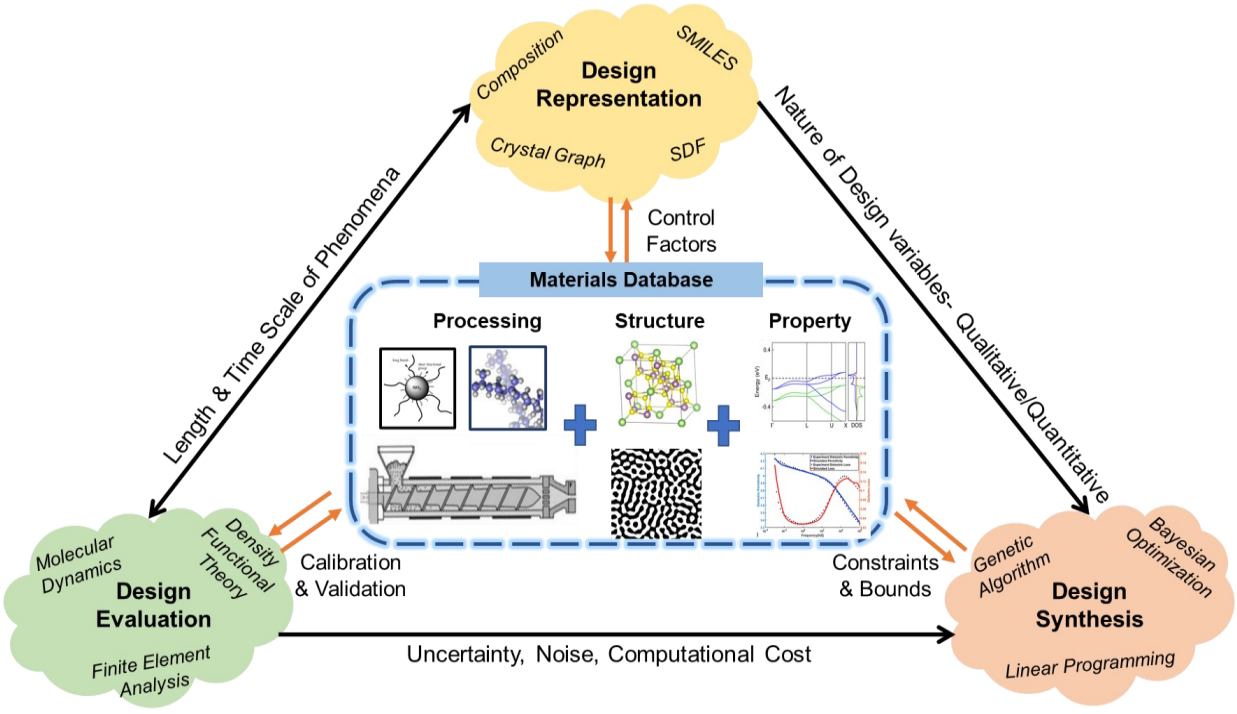


Figure 1-1: Data Centric Framework for Materials Design

Design Synthesis involves optimizing material properties to meet the requirements for specific applications. The choice of optimization method depends on nature of design variables – whether there are qualitative & quantitative design variables, presence of uncertainty/noise in property evaluations and its computational cost. To account for manufacturing feasibility and consistency with fundamental laws governing material properties, constraints and bounds are imposed during optimization to ensure feasible solutions.

1.2 Challenges in Data Centric Design

While existing methods have been successfully applied for some material systems, extending these techniques for wider applicability is fraught with challenges. We underline some these challenges below and categorize them under three outstanding themes in field of design engineering.

- **Design Representation**: Due to the high dimensionality of material microstructure, design (microstructure) representation is critical to ensure tractable design strategies. A good microstructure representation will (a) provide significant dimension reduction, (b) embody salient morphological features and (c) provide a computationally efficient reconstruction procedure. Given the vast diversity of microstructures observed in engineered products, developing an MCR technique that is universally applicable is challenging. Existing methods such as Correlation Functions, Physical Descriptors, SDF etc. are well suited for some systems while not for other, as examined in Bostanabad et al. [8]. However, we notice that there's a lack of methods suitable for representing anisotropic microstructures. Anisotropy is highly desirable in materials whose performance depends on an underlying transport phenomenon such as Thermoelectric devices, batteries, water filtration

membranes etc. Thus, the capability to characterize and reconstruct anisotropic microstructures is of primary interest. Characterization and reconstruction of multiphase microstructures represents another longstanding challenge in this domain. Baniassadi et al. [9] employed two point correlations to study relative distributions of phase and developed a Monte Carlo method for reconstruction. The computational expense of this reconstruction procedure and the inability of correlation functions to characterize complex microstructures displaying heterogeneity and anisotropy calls for the development of more robust methods.

- **Design Evaluation**: Developing design evaluation techniques largely depends on the properties of interest as well as the time & length scale at which these properties can be measured. It is highly desirable to have computationally efficient tools since the iterative nature of optimization algorithms requires several structure-property evaluations. Recently, machine learning techniques have become popular surrogates for physics driven structure-property models. But the requirement of sufficient amounts of training data and their inability to extrapolate places higher emphasis for computational efficient physics-based models. For Organic Photovoltaics, the development of Structure-Property relationship is challenging due to the intermingling of several phenomena and has prohibited design of active layer microstructure. Accounting for uncertainties, emanating from various sources such as variations in design variables, lack of data, model inadequacy, stochasticity is material systems etc., represents another challenge in this domain.
- **Design Synthesis:** Design Synthesis refers to the process of exploring the design space and identifying the optimum i.e., optimization. Given the highly non-linear behavior of material

properties and the computationally expensive simulation associated with property evaluation, optimization algorithms suitable to tackle these challenges are desirable. Depending on the nature of design variables, materials design can be cast as continuous variables, mixed variable or discrete variable problems. Existing methodologies are well suited for continuous variable problems and require significant feature engineering for discrete variables. Feature engineering for property prediction is challenging since knowledge of all influential factors is often unknown. Thus, methods to tackle discrete & mixed variable inputs with minimal feature engineering are sought. In addition to meeting performance objectives, functional materials design is multiobjective in order to satisfy auxiliary criterion such as manufacturing feasibility, cost, durability etc.

1.3 Research Tasks and Accomplishments

The objectives of this dissertation are to address some of the above-mentioned challenges in data centric material design highlighted in the previous section. In particular, we focus on the following task:

1. **Isotropic and anisotropic microstructure Design using Spectral Density Function**: Inspired by the recent success of SDF, we seek to extend and enhance its capabilities. First, we examine its applicability for design of active layer morphology in Organic Photovoltaics. Complex interactions of several phenomena coupled with the lack of understanding regarding the influence of fabrication conditions and nanostructure morphology have been major barriers to realizing higher PCE. To this end, we propose a SDF based computational microstructure design framework for designing the active layer of P3HT:PCBM based OPVCs conforming to the bulk heterojunction (BHJ) architecture.

Previous efforts on microstructure mediated design using SDF mostly assume isotropy, which is not ideal for applications where material properties along specific directions must be tailored to meet performance requirements, such as those associated with transport phenomena. We propose an anisotropic microstructure design strategy that leverages SDF for rapid reconstruction of high resolution, two phase, isotropic or anisotropic microstructures in 2D and 3D. We demonstrate that SDF microstructure representation provides an intuitive method for quantifying anisotropy through a dimensionless scalar variable termed anisotropy index.

2. **Concurrent composition and microstructure design**: 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. To this end, 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 Latent Variable Gaussian Process.

3. **Multicriteria optimization for combinatorial composition design**: Electronic materials exhibiting phase transitions between metastable states (e.g., metal-insulator transition materials with abrupt electrical resistivity transformations) are challenging to decode. For these materials, conventional machine learning methods display limited predictive capability due to data scarcity and the absence of features impeding model training. We demonstrate a discovery strategy based on multi-objective Bayesian optimization to directly circumvent these bottlenecks by utilizing Latent Variable Gaussian processes combined with high-fidelity electronic structure calculations for validation in the chalcogenide lacunar spinel family. We directly and simultaneously learn phase stability and band gap tunability from chemical composition alone to efficiently discover all superior compositions on the design Pareto front. Previously unidentified electronic transitions also emerge from our featureless adaptive optimization engine.
4. **Descriptor assisted Bayesian Optimization for materials design**: While Task 2 and 3 exemplify the efficacy of LVGP based Bayesian Optimization for mixed variable problems, qualitative input(s) with a large number of levels pose unique challenges. First, in order to estimate latent variable for each level, the training dataset for LVGP must contain at least one observation for each level of every qualitative variable. This constraint consequently leads to larger training datasets (i.e., high computational cost) to initialize Bayesian Optimization. Second, the number of latent variables to be estimated in the LVGP model increases linearly with number of levels of qualitative input. This consequently

increases the computational cost of model fitting. To overcome these challenges, we propose descriptor augmented LVGP that utilizes material descriptors to (a) identify a subset of levels to be included in the training dataset used to initialize Bayesian optimization, (b) predict latent variables for unobserved levels and (c) explore the feasible levels in future iterations. Through a variety of examples, we showcase the ability of descriptor augmented LVGP to overcome the pitfalls of partial and imperfect descriptor knowledge encountered in materials design.

5. **User friendly microstructure analysis tools for Material Science Community**: To provide quick & easy access to well-known microstructure analysis methods, we have implemented eight user-friendly webtools in NanoMine – an open source data repository for nanocomposites community [4]. The webtools perform all computations in the NanoMine server and send an email notification to users after their submissions have been processed. Additionally, all webtools support multiple image file formats and give users the freedom to analyze a single or set of images simultaneously.

1.4 Dissertation Outline

The outline of this dissertation and the interconnected nature of research tasks is shown in Figure 1-2. After laying out the challenges and research objectives in Chapter 1, Chapter 2 will provide technical background on essential concepts and methodologies utilized in the following chapters. Then, each research task is discussed separately in Chapters 3 through 6. A concise description of MCR webtools developed for NanoMine is provided in Appendix 9.1. The dissertation concludes in Chapter 7, first with a list of contributions (7.1) and then some future research themes identified by the author (7.2). Brief description of simulation methods used for

material property evaluation is provided in Appendix, along with copies of journal/news highlighting the work accomplished in this dissertation.

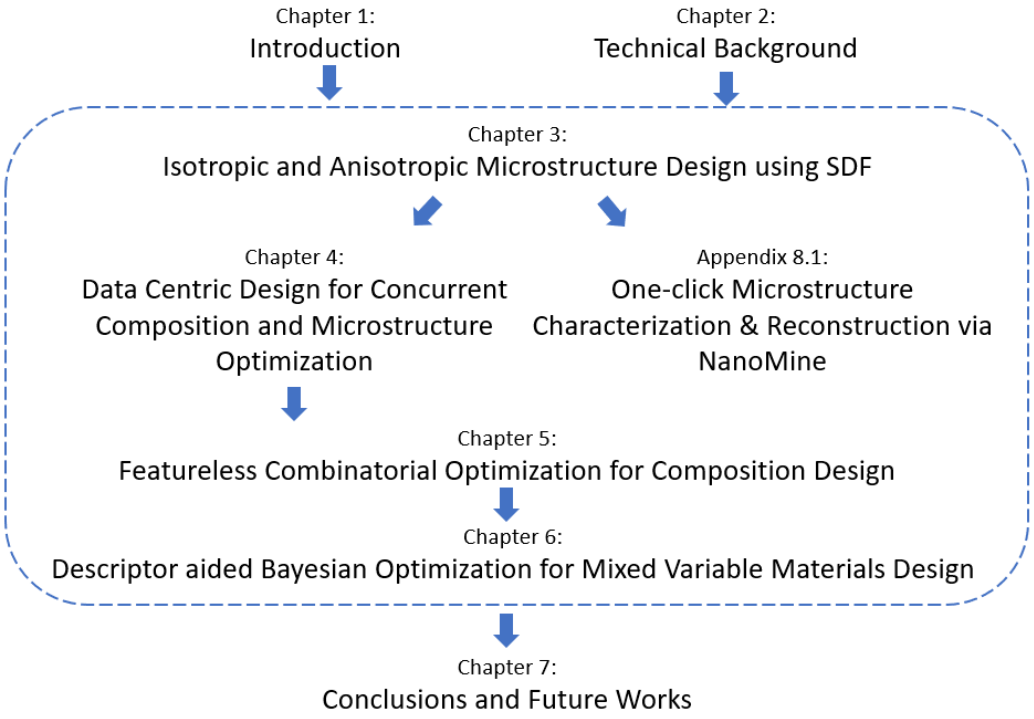


Figure 1-2: Outline of the dissertation

2 Technical Background

This chapter provides the technical background associated with the research tasks outlined in the dissertation.

2.1 Computational Microstructure Characterization and Reconstruction

As a material's morphology significantly influences its properties [10, 11], an essential task in creation of PSP linkages is analyzing the microstructure(s). The analysis is quantitative, and its outcome is a deep understanding of how processing conditions influence the formation of microstructure and how the microstructure in turn affects the properties. To analyze microstructures and extract useful morphological features, the following three-step strategy is recommended.

i. **Image binarization***:* Binarization is the process of converting a grayscale image to a black and white image (assuming there are only two phases – filler and matrix) by removing noise and consequently simplifying the analysis. The most widely used image binarization methods are Otsu's method [12] and the Niblack algorithm [13].

ii. **Microstructure Characterization***:* Several methods have been developed that can convert multi-dimensional microstructure morphology recorded in images into a set of functions (aka features/descriptors/predictors) that encode significant morphological details i.e., characterize the microstructure.

iii. **Microstructure Reconstruction:** After characterization, one can reconstruct a statistically equivalent microstructure(s) [14] which embodies a prescribed set of features (obtained by image characterization or provided by user) and can be used as a

representative volume element (RVE) for simulating material behavior via finite element analysis (which creates the structure-property linkage) or serve as a training dataset for machine learning algorithms [15, 16].

Microstructure Characterization and Reconstruction (MCR) approaches [17] for non-deterministic systems are based on spatial correlation functions [18-20], descriptor-based methods [21, 22] and machine learning techniques [16, 23, 24]. Some representative methods among these are summarized in Figure 2-1. Among the existing methods, Physical Descriptors [21, 22] and Spectral Density function [10, 25, 26] have been widely adopted for design of material systems due to their physically meaningful characterization, relative ease of reconstruction and low dimensional representation. The descriptor-based methodology aims to identify a set of uncorrelated features i.e., descriptors that represent salient morphological features. Reconstruction is accomplished through a hierarchical strategy to achieve microstructures with desired descriptors. Xu et.al. [22] used this method to design polymer nanocomposites for vehicle tire application. They used four descriptors for microstructure representation that also served as design variables – volume fraction, number of clusters, average elongation ratio and average nearest neighbor distance. Multiobjective Genetic algorithm was used to optimize viscoelastic properties, which resulted in identification of Pareto Front. Although descriptor based MCR provides greater control over morphological features, it is computationally expensive for reconstruction of high-resolution structures. Spectral Density Function (SDF) [10, 27-31], a frequency domain microstructure representation, has received a lot of attention for its capability to provide low dimensional, physically meaningful description of quasi-random material systems. For isotropic materials, SDF is one dimensional function of spatial frequency and represents spatial correlations

in the frequency domain. Although information contained in SDF is equivalent to two-point autocorrelation function, Yu et al. [27] have shown that SDF provides a more convenient representation for designing microstructures. However, all the above-mentioned studies relate to isotropic material systems and there are no instances of design of anisotropic microstructures to the best of our knowledge. This presents a major challenge since anisotropy is highly desired in some material systems, especially where the performance is a manifestation of an underlying transport phenomenon such as Organic Photovoltaic Cells (OPVCs).

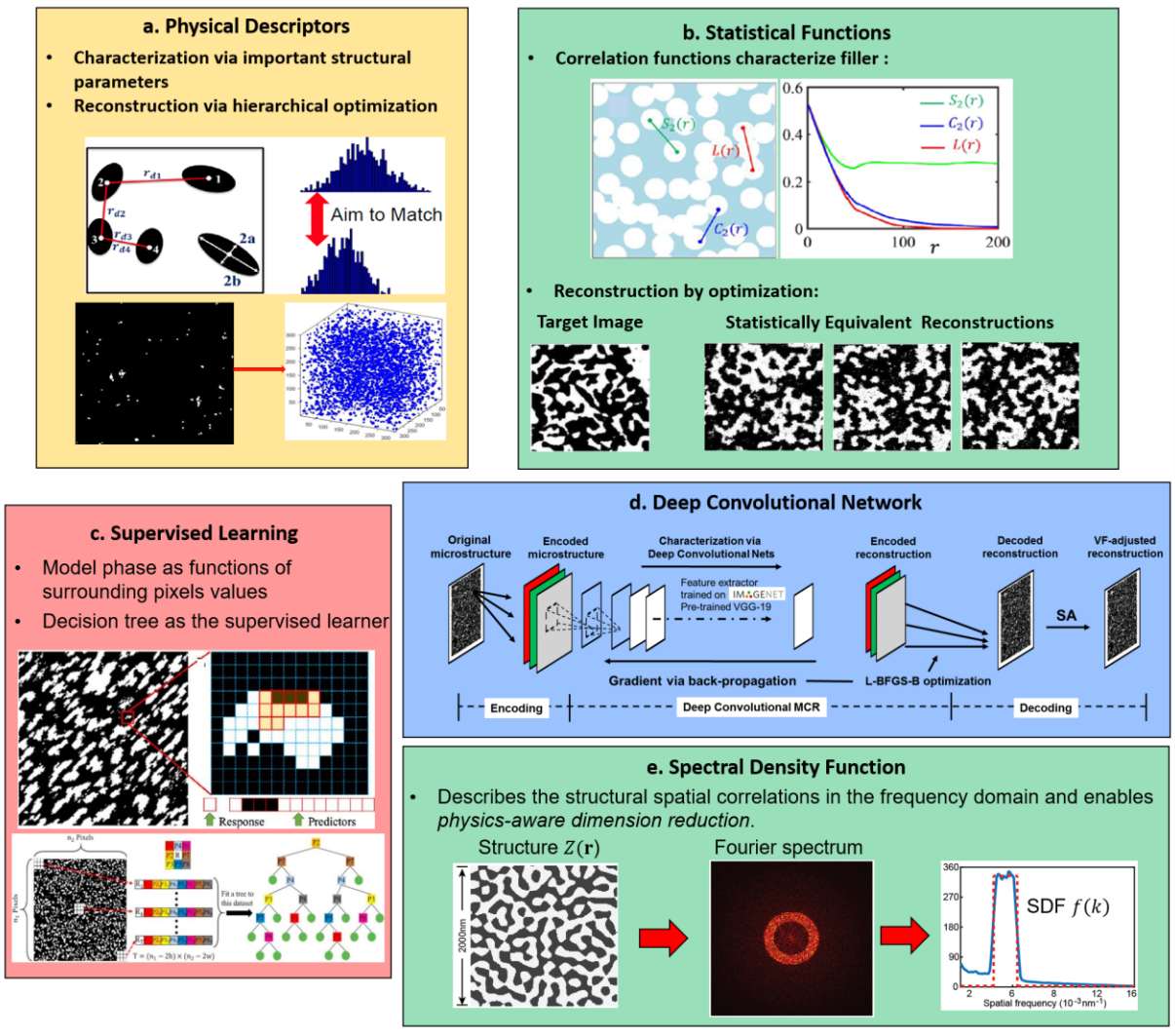


Figure 2-1: Representative microstructure characterization and reconstruction techniques

With the growing popularity of machine learning in the scientific community, several methods from this field have found ingenious new applications in MCR to overcome the assumptions (such as isotropy, stationarity) made in analytical methods. We categorize the application of machine learning methods for MCR in two categories: (a) reconstruction of a microstructure that's stochastically equivalent to a target and, (b) reconstruction of microstructures that mimic the distribution of a dataset - commonly referred to as generative modelling. The former category includes supervised learning [23, 32, 33], transfer learning [34, 35] and texture synthesis [36, 37] approaches. While these methods are extremely flexible and accurate, they do not provide an interpretable microstructure characterization and are not useful for microstructure design since they require a target microstructure. The latter category is largely dominated by generative models such as Variational autoencoders [38] and Generative Adversarial Networks (GAN) [39]. These methods are an attractive proposition for microstructure design tasks since they can be used to generate previously unseen microstructure and thus explore the space of feasible microstructures. Recent advancements in this field has seen the development of new methods that allow incorporation of processing parameters in the reconstruction process by conditioning the generator network of GANs [40, 41], thus addressing the processing-to-structure mapping that has been elusive in previous MCR models. However, the need of a large dataset to reliably train these generative models, which often include convolutional neural networks with several thousand parameters, remains a barrier to their universal adoption in microstructural design.

2.2 Spectral Density Function

The Spectral Density Function (SDF) (aka, Fourier power spectrum) is a low-dimensional representation of microstructure in the frequency domain where different frequencies represent

real space features at different length scales. It can be evaluated simply as the squared magnitude of the Fourier transform (FT) of a binary microstructure image $\boldsymbol{\mathcal{M}}$:

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

where $\mathcal{F}[.]$ denotes the Fourier transform operator and $\boldsymbol{k}$ is the frequency vector. Figure 2-2 depicts three isotropic, quasi-random channel-type nanostructures with ring shaped SDF. Channel-type nanostructures originate from bottom-up processes such as phase separation [42] or thin film wrinkling [43]. Figure 2-2(a) contains a single dominant frequency i.e., a single ring and manifests in channels with uniform width and connectivity. The channel width is inversely proportional to ring radius. Figure 2-2 (b,c) have additional rings at lower frequencies leading to wider channels with variations in channel width and increased disorder in nanostructure. Note that the type of nanostructure (and the form of SDF) is dependent on fabrication methods and materials used.

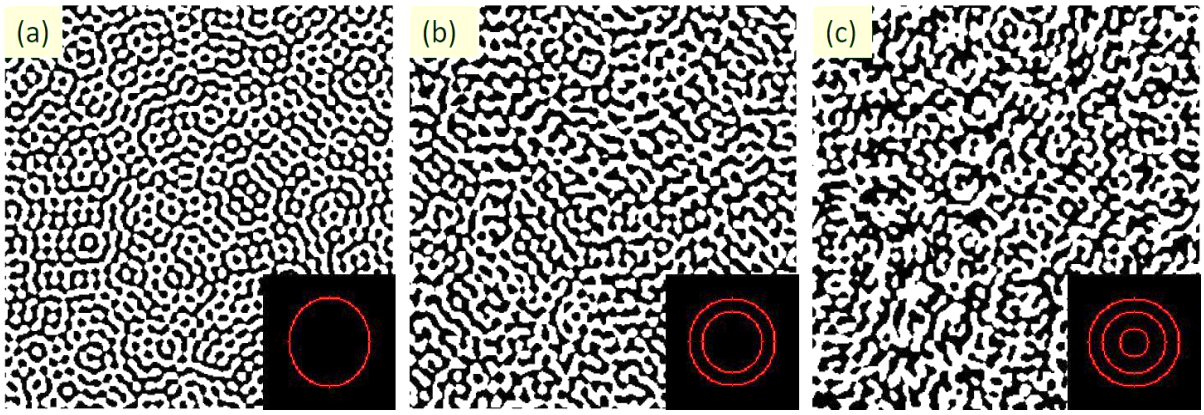


Figure 2-2: Three quasi-random nanostructures and their corresponding SDF shown in inset.

For isotropic microstructure, radial averaging can be used to convert vector $\boldsymbol{k}$ to a scalar (like radial averaging of position vector for correlation functions). According to the Winner-Khinchin Theorem [44], the inverse FT of SDF is the two-point autocorrelation function. Previous research suggests that SDF is sufficient to represent some complex heterogeneous microstructures with

irregular geometries. Studies have also shown that SDF is a physics-aware MCR technique that can map the SDF parameters to properties which largely depend on the spatial correlations of microstructures, for example optical properties [45].

Reconstructing microstructures with specified SDF can be accomplished in two distinct ways. Chen et.al [46] extended the Yeong-Torquato simulated annealing approach, first proposed for reconstruction using correlation functions, to construct disordered hyperuniform materials. Since this method uses pixel swapping strategy, it is computationally intensive. An alternative reconstruction approach assumes microstructures are realizations of a gaussian random field and aims to construct the field. One popular method for accomplishing this is Cahn's Scheme [47], an analytical approach for generating random fields with pre-specified SDF through

$$Y(\boldsymbol{r}) = \left(\frac{2}{N}\right)^{1/2} \sum_{i=1}^{N} \cos\left(k_i \widehat{\boldsymbol{k}}_i . \boldsymbol{r} + \varphi_i\right), \tag{2-2}$$

where φ_i and $\widehat{\boldsymbol{k}}_i$ are uniformly distributed on, respectively, $[0,2\pi]$ and a unit circle. k_i is a random variable distributed according to $P(k) = \rho(k)k$. N, the number of terms used in summation, is set to a large value (e.g., 10000) to ensure sufficient accuracy. The generated random field $Y(\boldsymbol{r})$ is then level-cut to obtain binary microstructures with desired volume fraction.

2.3 Latent Variable Gaussian Process

The standard GP methods were developed under the premise that all input variables are quantitative, which does not hold in many real engineering applications. We recently proposed a Latent Variable Gaussian processes (LVGP) [48] modeling method that maps the levels of the qualitative factor(s) to a set of numerical values for some underlying latent unobservable

quantitative variable(s) as illustrated in Figure 2-3. The 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 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.

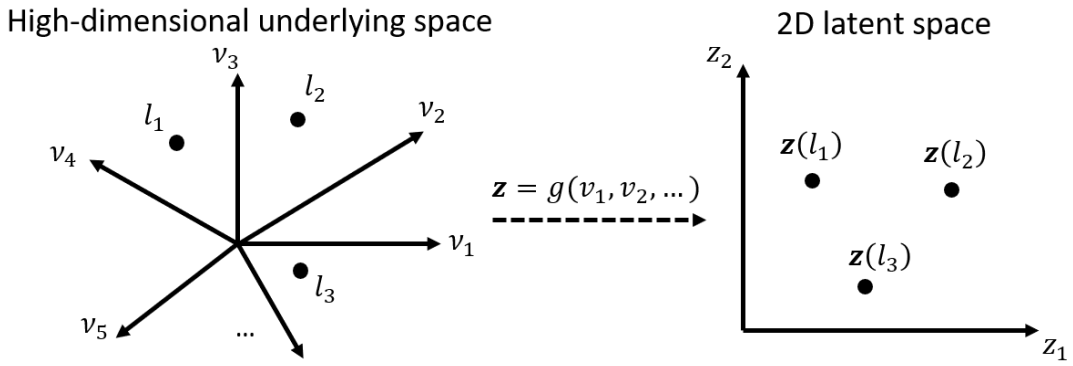


Figure 2-3: Illustration of high-dimensional underlying space of an arbitrary qualitative factor and the mapped latent space. The factor has levels l_1, l_2, and l_3, and is fully characterized by physical attributes v_1, v_2,…. The mapping $g: \boldsymbol{v} \rightarrow \boldsymbol{z}$ is implicitly constructed and found during the estimation of the latent variable values $\{\boldsymbol{z}(l_1), \boldsymbol{z}(l_2), \boldsymbol{z}(l_3)\}$.

A GP model (Y) can be represented as

$$y(\cdot) = \mu + G(\cdot), \tag{2-3}$$

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 $\boldsymbol{\varphi}$. In the LVGP method, the m 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 $\boldsymbol{z}^{(i)} \in \mathbb{R}^d$, where d is the dimensionality of $\boldsymbol{z}^{(i)}$. A modeler is 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 can be subsequently constructed as

$$r(\boldsymbol{w}, \boldsymbol{w}'|\boldsymbol{\varphi}, \boldsymbol{Z}) = exp\left\{-\sum_{i=1}^{p} \varphi_i\left(x_i - x_i'\right)^2 - \sum_{i=1}^{q} \left\|\boldsymbol{z}^{(i)}(t_i) - \boldsymbol{z}^{(i)}\left(t_i'\right)\right\|_2^2\right\}, \quad (2\text{-}4)$$

where $\boldsymbol{Z}$ is the collection of all the latent parameters denoted by $\{\boldsymbol{z}^{(1)}\left(l_1^{(1)}\right), \dots, \boldsymbol{z}^{(1)}\left(l_{m_1}^{(1)}\right), \boldsymbol{z}^{(2)}\left(l_1^{(2)}\right), \dots, \boldsymbol{z}^{(q)}\left(l_{m_q}^{(q)}\right)\}$. With this correlation formulation, hyperparameters $\boldsymbol{\varphi}, \boldsymbol{Z}, \mu$ & σ^2 can be found by maximizing the log-likelihood function ($\mathcal{L}$):

$$\mathcal{L}(\mu, \sigma^2, \boldsymbol{\varphi}, \boldsymbol{Z}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln|r(\boldsymbol{\varphi}, \boldsymbol{Z})| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mu\mathbf{1})^T r^{-1}(\boldsymbol{\varphi}, \boldsymbol{Z})(\boldsymbol{y} - \mu\mathbf{1}), \quad (2\text{-}5)$$

where n is the dataset size, $\mathbf{1}$ is n-by-1 vector of ones, $\mathbf{Y}$ is a n-by-1 vector of observed responses. Under this formulation of likelihood function, the closed form solution of μ and σ^2 can be found as a function of correlation matrix r which is subsequently a function of $\boldsymbol{\varphi}, \boldsymbol{Z}$ per Eq. (2-4):

$$\hat{\mu} = (\mathbf{1}^T r^{-1}\mathbf{1})^{-1}\mathbf{1}^T r^{-1}\boldsymbol{y}, \quad (2\text{-}6)$$

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \hat{\mu}\mathbf{1})^T r^{-1}(\boldsymbol{y} - \hat{\mu}\mathbf{1}), \quad (2\text{-}7)$$

Plugging the above equations in Eq. (2-5) results in a log-likelihood function which is solely a function of $\boldsymbol{\varphi}, \boldsymbol{Z}$**.** Subsequently, the optimization task devolves to identifying the optimum values of $\boldsymbol{\varphi}, \boldsymbol{Z}$**.**

After the optimal values of $\boldsymbol{\varphi}, \boldsymbol{Z}$ have been estimated (referred to as maximum likelihood estimates)**,** the response $\hat{y}$ at any new input $\boldsymbol{w}^*$ can be found as:

$$\hat{y}(\boldsymbol{w}^*) = \hat{\mu} + r(\boldsymbol{w}^*, \boldsymbol{w}) r^{-1}(\boldsymbol{w}, \boldsymbol{w})(\boldsymbol{y} - \hat{\mu}\mathbf{1}), \tag{2-8}$$

where $r(\boldsymbol{w}^*, \boldsymbol{w})$ is a matrix of pairwise correlations between $\boldsymbol{w}^*$ and each of the n observations in training dataset. Additionally, it is desirable to quantify the uncertainty associated with the prediction in Eq. (2-8) via predictive variance:

$$\hat{s}^2(\boldsymbol{w}^*) = \hat{\sigma}^2\{r(\boldsymbol{w}^*, \boldsymbol{w}^*) - r(\boldsymbol{w}^*, \boldsymbol{w}) r^{-1}(\boldsymbol{w}, \boldsymbol{w}) r(\boldsymbol{w}, \boldsymbol{w}^*)\}, \tag{2-9}$$

Unlike most supervised machine learning methods, LVGP does not require hand-crafted features to describe qualitative variables. Rather, it learns the underlying "latent variables" $(\boldsymbol{Z})$ influencing response $(\boldsymbol{y})$ by minimizing error in its prediction. Alleviating the need of feature engineering makes LVGP attractive for materials design applications.

2.4 Bayesian Optimization

Optimization is pervasive in academic and industrial settings since a wide variety of tasks can be cast as optimization problems. As pointed out in Figure 1, optimization is an eminent topic of research within design synthesis since all endeavors in engineering design strive to identify the optimal combination of design variables $\boldsymbol{w}_*$ out of all possible combinations contained in design space Ψ such that it minimizes a predefined yet unknown objective function f such that:

$$\boldsymbol{w}_* = \arg\min_{w \in \Psi} f(\boldsymbol{w}), \tag{2-10}$$

Among the many global optimization methods reported in literature, BO stands out due to its capability of locating the global optima for highly non-linear functions within tens of objective function (i.e., material property) evaluations. BO accomplishes this by repeating these three steps (illustrated in Figure 2-4)-

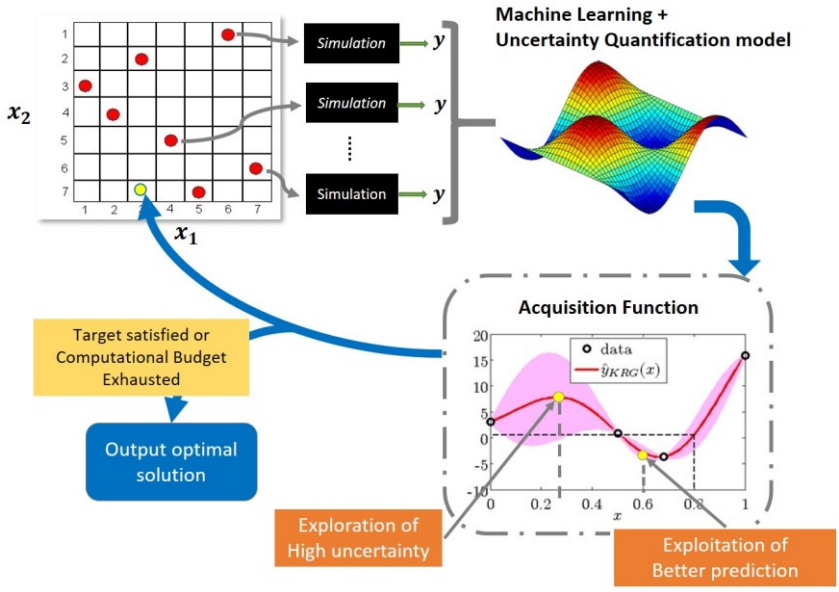


Figure 2-4: Bayesian Optimization framework

I. A machine learning model is trained on available data to predict material property ($\boldsymbol{y}$) of interest from the design variables $\boldsymbol{w} = (x_1, .. x_p, t_1, .. t_q)$ and supply uncertainty quantification over the design space.

II. An acquisition function uses the prediction and associated uncertainty to determine the best design to evaluate next. The acquisition function aims for exploration, exploitation, or both of the design space.

III. The design recommended by acquisition function is evaluated and added to the dataset.

This procedure is usually terminated after user-specified maximum iterations are completed. Gaussian Process modelling [49, 50] is a popular choice of ML model for BO due to its inherent ability for uncertainty quantification without incurring additional computational cost, although Random Forrest [51] and ensembles of Support Vector Machines [52] have also been used in the past. The acquisition function essentially gauges the benefit of evaluating a design by interrogating the ML model predictions and associated uncertainties. The acquisition function must decide between exploration and exploitation of design space, which may be contradictory goals. The best performing acquisition functions generally strike a balance between the two. Commonly used acquisition functions are probability of improvement (PI) [53] and expected improvement (EI) [54]:

$$PI(\boldsymbol{w}^*) = \Phi\left(\frac{\tau - \hat{y}(\boldsymbol{w}^*)}{\hat{s}(\boldsymbol{w}^*)}\right), \tag{2-11}$$

$$EI(\boldsymbol{w}^*) = \left(\tau - \hat{y}(\boldsymbol{w}^*)\right)\Phi\left(\frac{\tau - \hat{y}(\boldsymbol{w}^*)}{\hat{s}(\boldsymbol{w}^*)}\right) + \hat{\sigma}(\boldsymbol{w}^*)\phi\left(\frac{\tau - \hat{y}(\boldsymbol{w}^*)}{\hat{s}(\boldsymbol{w}^*)}\right), \tag{2-12}$$

where τ is minimum objective value observed so far, $\Phi(.)$ and $\phi(.)$ are the cumulative density function and probability density function of the standard normal distribution, respectively. The review article by Shahriari et al. [55] provides detailed discussions about these and other acquisitions functions used in BO.

Note that $\hat{y}$ and $\hat{s}$ are the predictive mean and variance at an unobserved design $\boldsymbol{w}^*$and are obtained from a statistical model such as Gaussian Process (described in Eq. (2-8) & Eq. (2-9)), Random Forrest etc.

3 Isotropic and Anisotropic Microstructure Design using Spectral Density Function

As noted in Chapter 2, the physics aware dimensionality reduction enabled by SDF has led to its adoption for microstructure design for a variety of material systems. However, all previous applications of SDF assume isotropy i.e., the salient microstructural features are identical across all directions. Consequently, the existing MCR capabilities of SDF fall short vis-à-vis anisotropic microstructures. Thus, there is a need to investigate the capabilities of SDF to characterize anisotropic microstructures and to develop a methodology for reconstruction and design. To this end, we seek to extend and enhance capabilities of SDF in this chapter. The challenge of designing high efficiency Organic Photovoltaics is used to demonstrate the efficacy of SDF based microstructure design. The major contributions discussed in this chapter are:

- We implement SDF based microstructure design framework for the Organic Photovoltaics using a novel performance evaluation methodology to identify optimal, isotopic active layer.
- Going beyond the traditional isotropic microstructure design formulations, we address the challenge of designing anisotropic microstructures in a computationally efficient manner. We develop a novel approach for fast microstructure reconstruction and quantify anisotropy to facilitate design of anisotropic microstructure.

3.1 Organic Photovoltaics (OPVC)

OPVCs are promising alternatives to traditional Silicon based solar cells due to several advantages – lightweight, flexibility, low production cost, short payback period [56, 57] etc. but,

their large-scale production and commercial usage has been plagued by problems of instability and batch to batch variability [58]. A typical OPVC shown in Figure 3-1 [59] consists of an active layer sandwiched between electrodes. Among the various electron donor/acceptor combinations investigated previously, phenyl-C61-Butyric-Acid-Methyl Ester (PCBM) interspersed with poly(3-hexylthiophene-2,5-diyl) (P3HT) has been the "best seller" [60]**.** There are four key processes taking place in the active layer during energy conversion process: (a) Exciton generation by light absorption; (b) exciton diffusion to donor:acceptor interface; (c) separation of charges from excitons to create electrons and holes and (d) movement of charges to respective electrodes. Owing to short mean free path of excitons [61], active layers conforming of Bulk Heterojunction (BHJ) architecture is key to ensure high efficiency. The morphology of BHJ, which is controlled by the processing method and related parameters, is crucial in deciding the performance of device [62, 63]. To optimize performance, one would like to maximize the donor:acceptor interfacial area (conversely, minimize the distance an exciton will need to travel) and ensure that the charges can reach their respective electrodes by traversing distance shorter than their mean free path. We leverage low dimensional microstructure representation enabled by SDF to optimize active layer morphology.

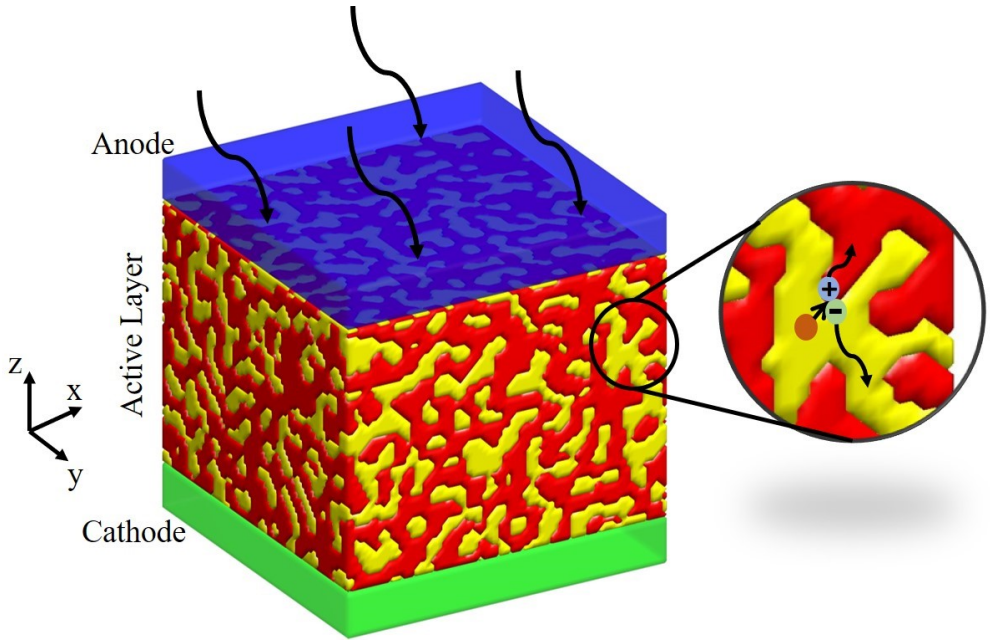


Figure 3-1: Schematic representation of Organic Photovoltaic cells and the energy conversion process. Magnified scene shows an exciton dissociating at the donor:acceptor interface, followed by charges migrating to respective electrodes.

3.2 SDF based Microstructure Design Framework for OPVC Active layer

Under the new paradigm of microstructure-sensitive material design [64, 65], materials are viewed as a complex structural systems that can be optimized for achieving superior properties (properties under consideration are subject to targeted application). Using OPVC active layer optimization as an example, we present here a holistic SDF based microstructure design framework (Figure 3-2) that can be employed for design of quasi-random nano- or microstructural systems based on structure-performance (S-P) relations [66].

The key idea of the proposed framework is to leverage SDF as the representation of OPVC microstructures, enabling direct and inverse S-P mappings. As shown in Figure 3-2, the framework is initiated by fabricating specimen of interest using a nanofabrication technique with processing

parameters choices based on empirical findings or literature. State-of-the-art imaging techniques are used to visualize the nanostructure in the available samples and the type (form) of SDF is identified. The main advantage of using SDF for quasi-random NMSs is that it can be easily parametrized and provides a more convenient representation for interpretation and design relative to other design methods [27]. Reconstruction is accomplished by level-cutting a Gaussian Random Field (GRF) governed by the required SDF. Thus, starting from a 2D XSTM/S image, SDF provides a reduced order microstructure representation (only three parameters required in this study) for creating statistically equivalent 3D microstructures which serve as Representative Volume Element (RVE) for performance evaluation.

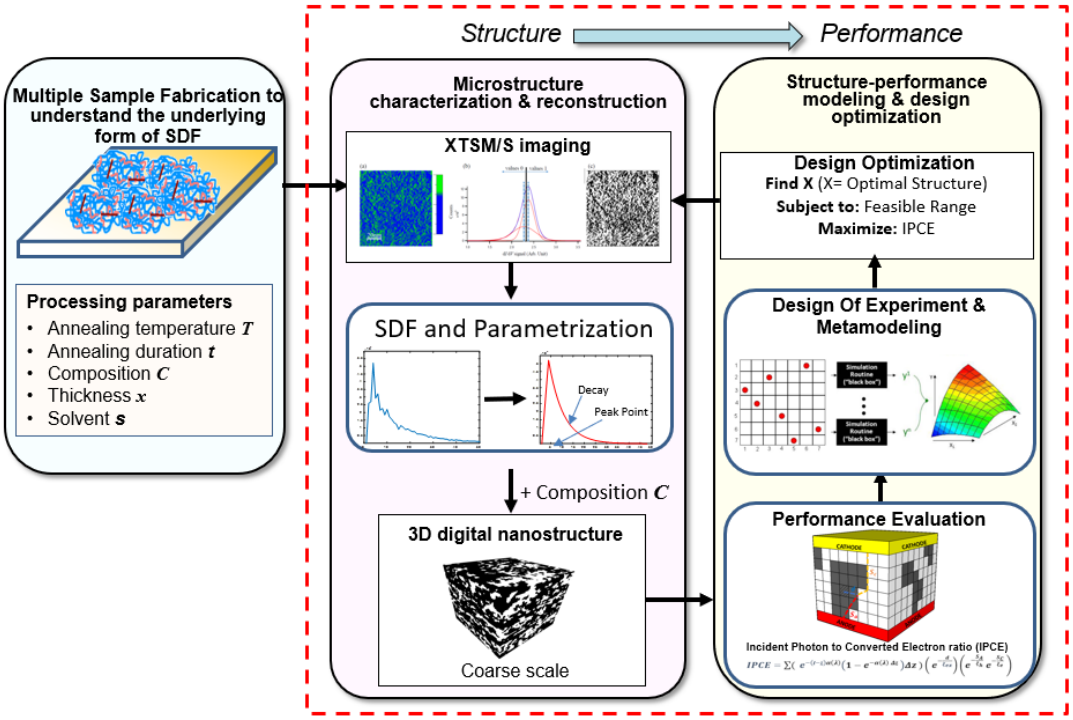


Figure 3-2: A framework for designing active layer nanostructure in bulk heterojunction OPVC via Spectral Density Function

To evaluate the performance of an RVE, we seek a model that accounts for structural features in addition to device physics and material properties. For OPVCs, the key performance parameter of interest is the Incident Photon to Converted Electron (IPCE) ratio. To evaluate IPCE computationally, a novel strategy based on device physics and nanostructure is developed here. This strategy explicitly states the influence of nanostructure on known physical phenomena and establishes the S-P relationship that forms the basis for performance optimization. However, before the optimization, creating a metamodel to replace the computationally expensive and time-consuming S-P model is highly desirable. Metamodel, created by careful Design of Experiments (DOE) [67], is essentially a "black-box" that approximates the S-P simulations. Given the set of design variables and their bounds, DOE dictates the S-P simulations that must be performed to determine the corresponding value of objective function. A suitable machine learning model is chosen to interpolate between known values of response, forming a metamodel which can be queried at each iteration of the optimization. In this study, we use Optimal Latin Hypercube Sampling (OLHS) to create the metamodel based on the Kriging method, accelerating the search for the optimal design.

Design optimization is performed with the pre-determined design variables obtained by parametrizing SDF along with the material composition. In this work, IPCE is chosen as the objective function with an aim of finding its maximum value and the corresponding SDF parameters (i.e., nanostructure). However, it should be noted that the optimum structure is limited to the same type of material system as the fabricated samples because the form of the SDF function used for optimization is determined based on the fabricated samples. In the following sections, we elaborate the procedure of implementing the proposed framework.

3.3 Structure-Performance Simulation Model

Here, we present an equation to predict the Incident Photon to Converted Electron (IPCE) from a 3D active layer microstructure. Under the finite element scheme, the equation for estimating performance from the microstructure can start with counting the number of collected electrons/holes per unit time through a summation of contributions from each volume element (voxel) in the active layer as:

$$\frac{n(\lambda)}{\Delta t} = \sum \left(\frac{I(\lambda)}{h\,\frac{c}{\lambda}}\, e^{-(t-z)\alpha(\lambda)} \Delta x \Delta y P_{ex}(\lambda) \right) \left(e^{-\frac{d}{\xi_{ex}}} \right) \left(P_{sep} \right) \left(e^{-\frac{S_A}{\xi_h}}\, e^{-\frac{S_C}{\xi_e}}\, P_{col} \right), \tag{3-1}$$

The four parentheses in Eq. (3-1) represent the four steps illustrated in Figure 3-1: (i) light absorption (exciton creation); (ii) Exciton diffusion; (iii) charge separation; and (iv) charge diffusion & collection. Here, t is the thickness of the active layer; $\alpha(\lambda)$ is the absorption coefficient of active layer as function of the light wavelength, λ; P refers to probability for exciton creation (ex), for charge separation (sep), and for charge collection (col); d is the distance to the nearest interface from the location of the exciton creation; ξ the diffusion lengths of exciton (ex); of hole (h); and of electron (e); S are the lengths of the path to anode (A); and to cathode (C). In this equation, the recombination behaviors of the charges are simply assumed to follow exponential decay over the distance it moves.

From previous study the value of ξ_e (diffusion length for electron) is found to be ~340 nm; the value of ξ_h (diffusion length for hole) is found to be ~90 nm [68]; the value of ξ_{ex} (diffusion length for exciton) is found to be 5.4 ± 0.7 nm [69] and $\alpha(\lambda)$ (absorption coefficient) is measured and could be found in [70].

Among the variables in Eq. (3-1), the probability of exciton creation, $P_{ex}(\lambda)$, and the absorption coefficient, $\alpha(\lambda)$, could be related to each other through the following relationship:

$$\mathrm{P_{ex}} = 1 - \mathrm{e}^{-\alpha(\lambda)\,\Delta z}, \tag{3-2}$$

This relationship is deduced by assuming:

$$\mathrm{P_{ex}}(\lambda) = \left(\frac{\text{no. of photon absorbed}}{\text{no. of photon incident}}\right) \cdot \left(\frac{\text{no. of exciton created}}{\text{no. of photon absorbed}}\right), \tag{3-3}$$

where the second term is closely related to the internal quantum efficiency, which is very close to 100 % in many cases [71], so it is assumed to be 1 here. On the other hand, the first term is closely related to the photon absorption coefficient, $I(z) = I_o\, e^{-\alpha(\lambda)z}$.

Thus Eq. (3-1) could be expressed as:

$$\frac{n(\lambda)}{\Delta t} = \sum \left(\frac{I(\lambda)}{h\frac{c}{\lambda}}\, e^{-(t-z)\alpha(\lambda)} \Delta x \Delta y \left(1 - \mathrm{e}^{-\alpha(\lambda)\,\Delta z}\right) \right) \left(e^{-\frac{d}{\xi_{ex}}} \right) \left(P_{sep}\right) \left(e^{-\frac{S_A}{\xi_h}}\, e^{-\frac{S_C}{\xi_e}}\, P_{col} \right), \tag{3-4}$$

We compute the IPCE, which is the number of electrons collected per incident photon, as: $IPCE(\lambda) = \frac{n(\lambda)}{\Delta t} \frac{1}{\frac{I(\lambda)}{h\frac{c}{\lambda}} A}$, where A represents the area of the sample illuminated by light. In this simulation, it is assumed that the whole sample surface is illuminated by light, indicating the sample surface area is A. The final working equation for evaluating IPCE from nanostructure is expressed as the summation over every voxel:

$$IPCE(\lambda) = \frac{1}{A} \sum \left(\left(e^{-(t-z)\alpha(\lambda)} \Delta x \Delta y \left(1 - e^{-\alpha(\lambda)\,\Delta z}\right) \right) \left(e^{-\frac{d}{\xi_{ex}}} \right) \left(P_{sep}\right) \left(e^{-\frac{S_A}{\xi_h}}\, e^{-\frac{S_C}{\xi_e}}\, P_{col} \right) \right), \tag{3-5}$$

where, z, d, $\mathrm{S_A}$, and $\mathrm{S_C}$ of each voxel are determined from the nanostructure. Eq.(3-5) is used to evaluate the performance of the active layer nanostructures in this study.

3.4 Isotropic Active layer Design

We leverage the low-dimensional structure representation enabled by SDF to formulate a design paradigm using a small set of variables. The active layer is represented by an RVE of size 80 x 80 x 80 voxels; equivalent to 100nm x 100nm x 100nm. Two SDF parameters–Peak Point and Decay – account for structural characteristics that control the charge separation and transport phenomenon discussed above. Further, the assumption that exciton generation is restricted to P3HT molecules necessitates the inclusion of donor/acceptor composition as an additional design variable. Here, PCBM volume fraction is chosen as the composition design variable. Note that composition also plays a critical role in level cutting the GRF for the reconstruction. Thus, only three variables are required: two from SDF plus the PCBM volume fraction.

The bounds for design variables are identified by analyzing the SDF of the two fabricated samples to estimate the three SDF parameters. Then a broad range for each of the three parameters is selected to ensure diverse SDF curves Previous studies, focusing only on active layer composition, have revealed that the ideal PCBM volume fraction (VF) is 0.37 approximately [72, 73]. To explore a wider range of values around the optimum, we allow VF to vary between 0.15 and 0.75. With the objective of maximizing IPCE ratio, the optimization problem can be stated as:

$$\max_{\mathbf{m}\in\mathbf{M}} IPCE \tag{3-6}$$

where **M** represents the set of all feasible microstructures characterized by p (peak point) $\in [2,10]$, d (decay) $\in [1,12]$, and vf (volume fraction) $\in [0.15,0.75]$.

Since optimization is an iterative process, it requires several S-P simulations (constructing RVE for current value of design variables and evaluating the IPCE ratio). The computational cost associated with reconstructing and evaluating a 80^3 voxels RVE is significant. To overcome this

computational burden and accelerate optimization, a metamodel is used. 45 OLHS [74] design were used for creating the Gaussian Process (GP) metamodel with three design variables while 11 were used for cross-validation. The R-Squared value based on validation points is 0.9792, which indicates a fair fit. Because of the highly nonlinear response of the metamodel, Genetic Algorithm(GA) is applied to obtain the global maximum IPCE. To test accuracy, multiple starting designs were selected. For all starting points considered in this study, the optimization routine converges to the design {Peak Point = 2, Decay = 12, VF = 0.2764 and IPCE = 8.41%}. This result relates to a 36.75% increase in IPCE ratio compared to experimental specimen which has an IPCE ratio of 6.15%. An RVE is reconstructed (Figure 3-3) using the optimal microstructure design variables and its IPCE ratio is computed. Compared to 8.41% from the metamodel, the reconstructed RVE results in an IPCE ratio of 8.19%, reinforcing the fact that the metamodel used here is sufficiently accurate.

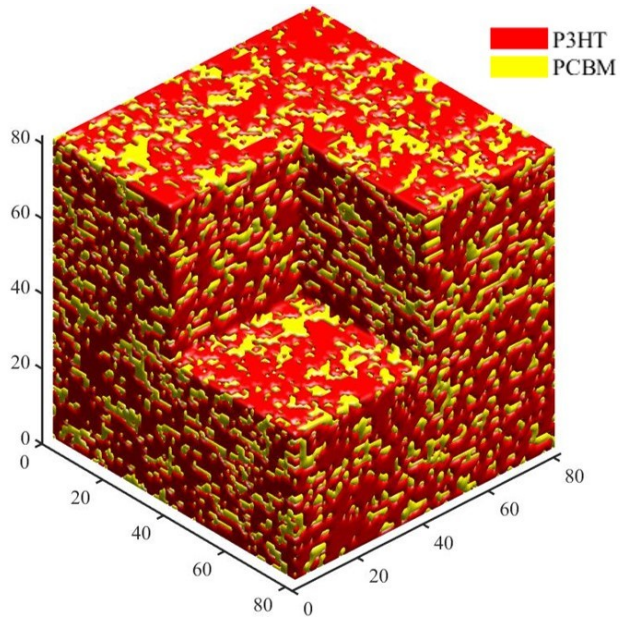


Figure 3-3: Optimal active layer microstructure

Although the optimized design enhanced P3HT:PCBM interfacial area; it does not represent optimal morphology w.r.t charge transport. This is because the tortuosity of PCBM domains forces electrons to traverse longer paths to reach cathode and in some instances, PCBM domains are isolated and do not provide any paths to cathode. Consequently, IPCE is diminished. Intuition dictates that orienting P3HT:PCBM domains in the direction of electrodes i.e., anisotropy will shorten the distance traversed by charges and enhance IPCE.

3.5 Reconstructing Anisotropic Microstructures using SDF

Although the optimized design enhanced P3HT:PCBM interfacial area; they do not represent optimal morphology w.r.t charge transport. This is because the tortuosity of PCBM domains forces electrons to traverse longer paths to reach cathode and in some instances, PCBM domains are isolated and do not provide any paths to cathode. Consequently, IPCE is diminished. Intuition dictates that orienting P3HT:PCBM domains in the direction of electrodes i.e., anisotropy will shorten the distance traversed by charges and enhance IPCE.

The subsequent sections in this chapter address the challenge of anisotropic microstructure design by presenting a novel SDF based reconstruction technique for rapid microstructure generation. Our method is based on Inverse Fourier Transform and can be implemented parsimoniously in any computational package. The method enables reconstruction of anisotropic microstructures without modification. Further, an SDF based anisotropy index is defined to quantify anisotropy and serve as an additional descriptor in the design of strongly anisotropic OPVC active layer that outperforms isotropic designs. We first describe Linear Time Invariant Systems, which is the foundation for reconstruction procedure and demonstrate its applicability to

isotropic/anisotropic microstructure reconstruction. Further, we propose anisotropy index as a quantitative measure of anisotropy and present two examples to underline its significance.

3.6 Linear Time Invariant Systems

Linear Time Invariant (LTI) systems [75] combine two useful concepts in digital signal processing namely linearity & time invariance. A system is linear when adding or scaling inputs to the system results in adding or scaling of outputs. Time invariance implies that a time delay in system inputs results in an equivalent delay in the outputs, without changing the system behavior. With these two essential properties, powerful analysis tools have been developed to study behavior of LTI systems. An important result in this domain is that an LTI system can be completely characterized by its impulse response, which is the output of system when input is a unit impulse. Once the impulse response is known, the system output (Y) to any arbitrary input (X) can be deduced using the relationship [76]:

$$\rho_Y(f) = |H(f)|^2 \rho_X(f), \tag{3-7}$$

where ρ_X, ρ_Y are the SDF of input and output signals, H is the Fourier Transform of system's impulse response and f denotes frequency. SDF (ρ) of a signal X is the squared magnitude of its Fourier Transform:

$$\rho(f) = |\mathcal{F}[X]|^2, \tag{3-8}$$

where $\mathcal{F}[.]$ represents the Fourier Transform operator.

The concept of LTI systems can be extended for cases where signals vary spatially rather than over time and generalized as Linear Shift Invariant (LSI) systems [77]. A microstructure can be considered a 2D signal and manipulated using LSI system as described below.

3.7 Fast Reconstruction using Spectral Density Function

SDF based MCR is best suited for quasi-random microstructures that exhibit a seemingly random material distribution but governed by an underlying correlation function. SDF represents spatial correlation in the spatial frequency domain, providing a simplified and physics aware representation of microstructure. A homogeneous microstructure can be described as a realization of an underlying stationary random field and reconstruction involves finding the random field with a prescribed SDF. Towards this end, we propose casting the reconstruction process as an LSI system that takes in a random white noise image and transform it into an image (microstructure) with desired SDF. This is accomplished by rewriting Eq. (3-7) as:

$$\rho_R(\boldsymbol{k}) = \rho_T(\boldsymbol{k}).\rho_W(\boldsymbol{k}), \tag{3-9}$$

where . represents point-wise multiplication, subscripts R, T & W denote the reconstructed, target & white noise SDF and $\boldsymbol{k}$ is a vector representing spatial frequency. Note that the white noise image and target image must have the same resolution. SDF is the squared magnitude of Fourier transform and hence, we can recover the reconstructed microstructure from Eq. (3-9) by level cutting $\mathcal{M}_R$ to the desired composition of two phases:

$$\mathcal{M}_R = \left|\mathcal{F}^{-1}\left\{\sqrt{\rho_T(\boldsymbol{k})}.\mathcal{F}\{\mathcal{M}_W\}\right\}\right| = \left|\mathcal{F}^{-1}\left\{|\mathcal{F}\{\mathcal{M}_T\}|.\mathcal{F}\{\mathcal{M}_W\}\right\}\right|, \tag{3-10}$$

Here, subscripts $\mathcal{M}_R, \mathcal{M}_T \; and \; \mathcal{M}_W$ denote reconstructed, target and white noise microstructures respectively. Since a white noise image contains all frequencies in equal measure, Eq. (3-10) can be interpreted as follows: the reconstruction process works like an LSI system with impulse response $\mathcal{M}_T$ to filter out all frequencies from white noise image $\mathcal{M}_W$ except the ones present in $\mathcal{M}_T$. Analogously, reconstruction is a convolution between a white noise image and

target image. Figure 3-4 shows some examples of microstructures generated using Eq. (3-10) by supplying a target SDF $\rho_T(\boldsymbol{k})$. Note that $\mathcal{M}_R$ will have the same resolution as $\mathcal{M}_T$ (or ρ_T). The use of white noise image in Eq. (3-10) introduces stochasticity, a key feature across all MCR methods and leads to an ensemble of statistically equivalent reconstructions sharing a common SDF as shown in Figure 3-4 (A & B). Since we have shifted the zero frequency component to the center of spectrum for target SDF (ρ_T) shown in insets, an identical zero frequency shifting must be performed on $\mathcal{F}\{\mathcal{M}_W\}$ prior to application of Eq. (3-10).

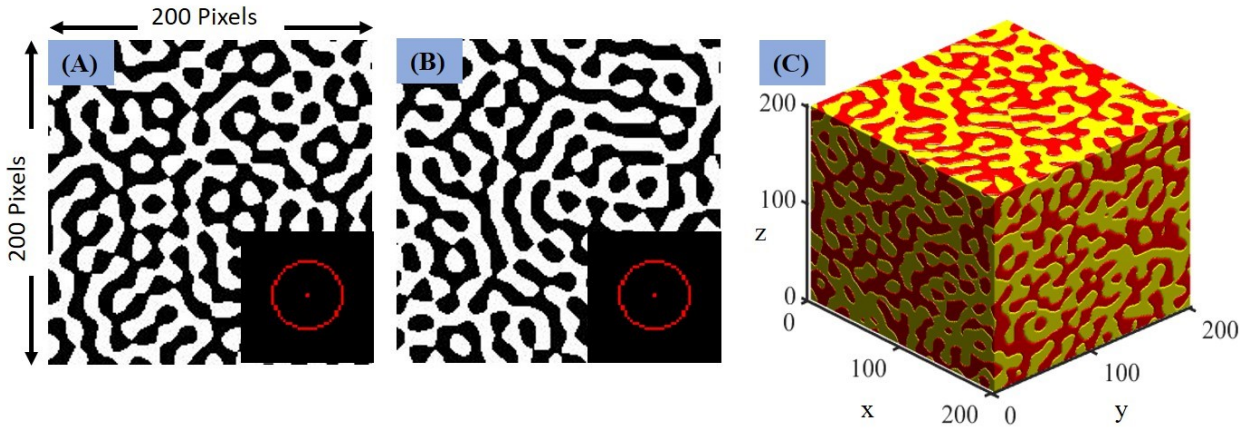


Figure 3-4: SDF based microstructure reconstruction. (A & B) Two 2D microstructures (200 x 200 pixels) generated from identical ring-type 2D SDFs shown in inset (zero frequency component shifted to center of spectrum). White phases volume fraction is 50% in both cases. In the plot of SDF, the black regions represent frequencies with zero intensity. (C) 3D microstructure (200 x 200 x 200 voxels) generated from an equivalent 3D ring type SDF. Yellow phase volume fraction is 50%.

SDF based reconstruction using Eq. (3-10) has two primary advantages as compared to existing methods. First, the reconstruction process only involves Fourier transform and Inverse Fourier Transform, which can be accomplished very efficiently using Fast Fourier Transform available in

any computational software package. This is highly significant in terms of generating high resolution 3D microstructures such as those required for investigating OPVCs and performance optimization which will require evaluation of structure-performance model at each iteration. Table 3-1 compares Cahn's method against the proposed method w.r.t computational time required for 2D and 3D reconstructions. These tests were performed on a 3.60GHz Intel ® Core™ i7 processor with 24GB RAM. It is quite evident that our method is significantly more efficient than existing methods.

Table 3-1: Examining computational efficiency of reconstruction methods.

Dimension	**Resolution**	**Computational Time**	
		Our Method	Cahn's Method
2D	100 x 100 pixels	0.002 seconds	0.175 seconds
	200 x 200 pixels	0.003 seconds	0.510 seconds
	400 x 400 pixels	0.008 seconds	2.254 seconds
3D	50 x 50 x 50 voxels	0.039 seconds	176.511 seconds
	100 x 100 x 100 voxels	0.087 seconds	1291.499 seconds
	200 x 200 x 200 voxels	0.802 seconds	3.3 hours
	400 x 400 x 400 voxels	7.337 seconds	Out of Memory

The second advantage of the proposed method is that no modifications are required for reconstruction of anisotropic microstructures using Eq. (3-10) since it is valid for any homogeneous, two-phase microstructure. This is elaborated in following section.

3.8 Spectral Density Function based Anisotropy Index

Unlike isotropy, which is an absolute state, anisotropy is a relative state and necessitates an appropriate quantitative measure. For example, Figure 3-4(A) & Figure 3-5(A1) are isotropic

microstructures and there is no meaningful notion of one being more isotropic than the other. On the other hand, Figure 3-5(B1 & C1) are anisotropic but, Figure 3-5 (C1) is more anisotropic than Figure 3-5 (B1). This example highlights the need for a metric which captures the degree of anisotropy in microstructure. We define an SDF based anisotropy index with the observation that dominant structural features in spatial domain manifest as non-zero values of frequency components. The pattern formed by these dominant frequencies depends on the microstructural features. Here we discuss two such patterns – ring and disk type SDFs but the concept can be easily generalized to other patterns.

For ring type SDF, we define anisotropy index α as the sine of polar angle ω subtended by the non-zero frequency component on the axis of anisotropy.

$$\alpha = \sin(\omega), \tag{3-11}$$

Figure 3-5 shows three sample 2D & 3D microstructures with different degrees of anisotropy, quantified by α. As noted earlier, the dimension of SDF is same as that of microstructure. Thus, the 2D microstructures (Figure 3-5 A1,B1,C1) have 2D SDFs while 3D microstructures (Figure 3-5 A3,B3,C3) have 3D SDFs shown in Figure 3-5 A2,B2,C2. Microstructures were generated by supplying the corresponding SDFs as $\rho_T(\boldsymbol{k})$ in Eq. (3-10). We observe that SDF's of isotropic microstructures possess symmetry about the center (zero frequency) while anisotropy arises from deteriorating symmetry.

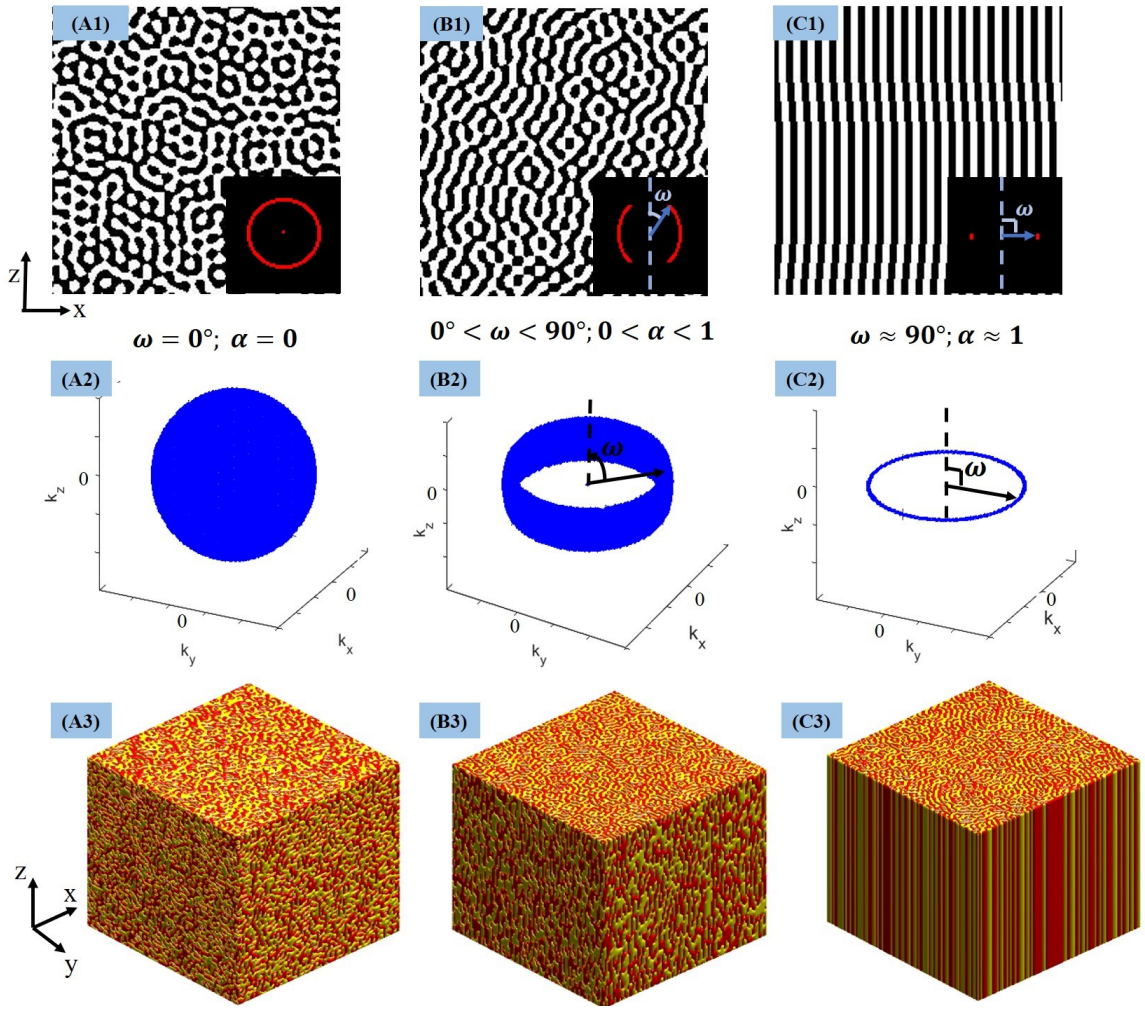


Figure 3-5: (A1,B1,C1) 2D microstructures with varying degree of anistropy and their SDFs shown in inset. (A3,B3,C3) 3D microstructures and their corresponding SDFs (A2,B2,C2). A1-A3, B1-B3 and C1-C3 represents isotropic, anisotropic and strongly ansiotropic microstructures. Volume fraction of each phase is 50% in all 2D and 3D microstructures shown above.

The angle ω captures the extent of symmetry deterioration and thus forms a suitable measure. Eq. (3-11) bounds α in the interval [0,1], where 0 and 1 represent isotropy and absolute anisotropy

respectively and facilitates its usage as a bounded design variable in the optimization case study presented in the following section.

Anisotropy Index is, in general, a measure of skewness in SDF pattern and can be extended beyond ring type SDFs shown in Figure 3-6 Consider the 2D microstructures in Figure 3-6 which exhibit disk type SDF patterns. Since there are multiple frequencies present in these microstructures, they appear more disordered than those in Figure 3-6 which have only one frequency present. A convenient measure of anisotropy i.e., anisotropy index α for these microstructures would be the eccentricity of SDF patterns. Eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. Its value ranges from 0 (isotropic) to 1 (strongly anisotropic). These microstructures were reconstructed from Eq. (3-10) by specifying the SDFs shown in inset as $\rho_T(\boldsymbol{k})$. Thus, our method generates anisotropic microstructures when $\rho_T(\boldsymbol{k})$ is anisotropic; without any modification of Eq. (3-10).

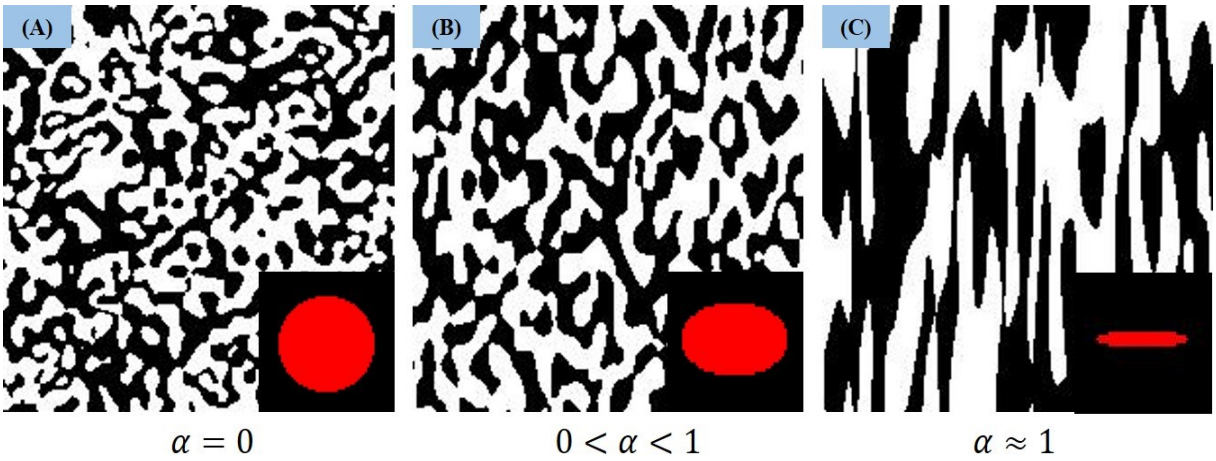


Figure 3-6: Quantifying anisotropy for micorstructures with elliptical SDF. Each microstructure is 200 x 200 pixels with 50% white phase area fraction. Inset shows corresponding SDFs. Anisotropy index α is defined as eccentricity of SDF pattern.

3.9 Revisiting OPVC Design Case Study

As noted earlier, anisotropic active layer can enhance OPVC performance but, realizing such designs has been a challenge hitherto. Here we demonstrate how the SDF based anisotropic microstructure design strategy addresses this challenge.

Given the vast space of possible active layer morphologies, microstructure optimization is necessary to identify active layer designs that maximize IPCE ratio. The IPCE ratio encapsulates the efficiencies of charge conversion and transport processes which are strongly influenced by active layer microstructure. The procedure to evaluate IPCE is provided in Appendix. Ideally, we desire an active layer that maximizes interfacial area and provides short, well-connected pathways for charge transport; but realizing both objectives simultaneously is a challenging task. The low-dimensional microstructure representation enabled by SDF is leveraged to formulate the active layer design as follows:

$$\max_{\mathrm{m}\in\mathcal{M}} \ \mathrm{IPCE}, \tag{3-12}$$

$$\boldsymbol{\mathcal{M}}: microstructures\ with\ 0\% \le VF_{PCBM} \le 100\%, 0.01nm^{-1} \le k_i \le 2.23nm^{-1}, 0 \le \alpha \le 1$$

where $\mathcal{M}$ is the set of all feasible microstructures, VF_{PCBM} denotes the PCBM concentration by volume ($VF_{PCBM} + VF_{P3HT} = 1$). Here we assume SDF follows a ring type pattern with radius k_i, ring thickness $0.01nm^{-1}$ and anisotropy index α. k_i controls width of PCBM domains; large values lead to narrower PCBM domains and vice-versa. Bayesian Optimization [78, 79], which adaptively samples designs to efficiently identify global optimum, was applied to solve the formulation presented in Eq. (3-12). In each iteration, a high-resolution 3D microstructure (450 x 450 x 450 voxels), embodying a 100nm x 100nm x 100nm Representative Volume Element, is generated using the SDF based reconstruction method discussed previously and its IPCE is

evaluated. To show the benefits of anisotropy, we also present results from optimization of isotropic microstructure with design variables VF_{PCBM} and k_i bounded in the range specified by Eq. (3-12) but α fixed to 0. We performed 100 iterations of Bayesian optimization with expected improvement [54] acquisition criterion.

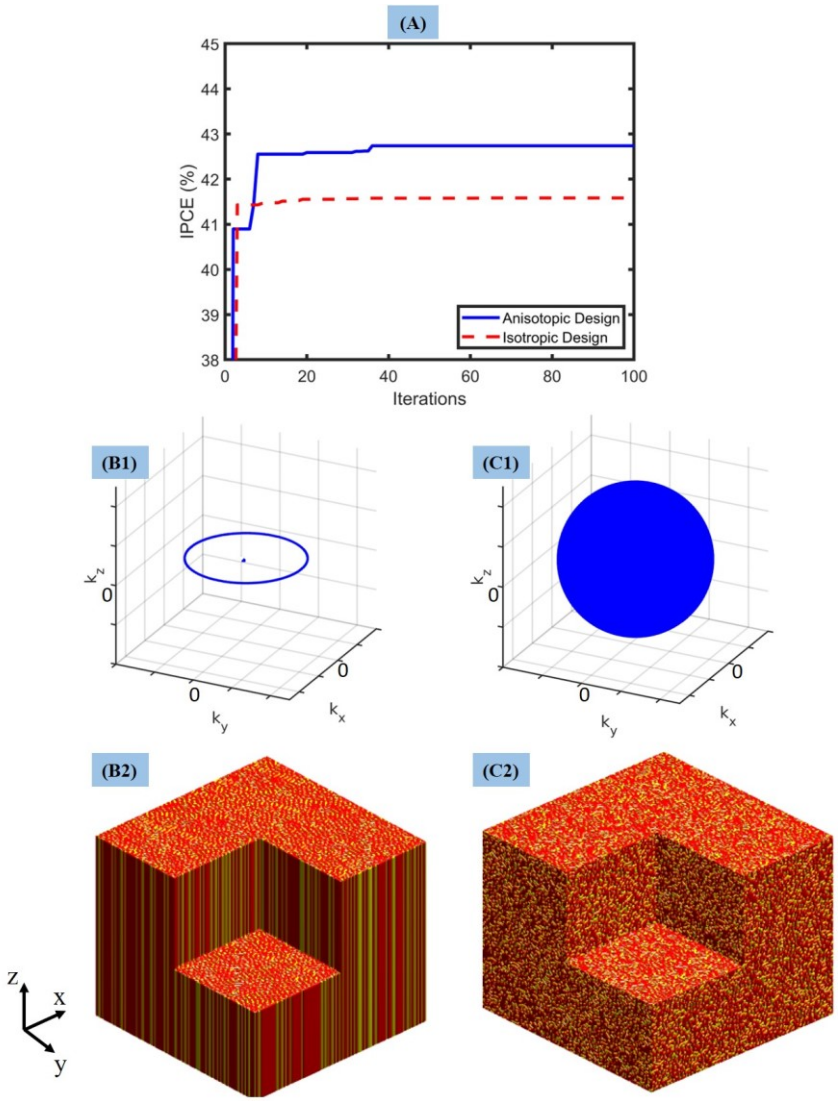


Figure 3-7: (A) Bayesian Optimization history for IPCE maximization, depicting superior perfomance of anisotropic design as compared to isotropic design. (B1) SDF of Optimized anistropic microstructure (B2) displaying perfect anisotropy along Z – direction. (C1) SDF of

Optimized istropic microstructure (C2). Red and yellow phase represents P3HT & PCBM respectively.

Figure 3-7A indicates that anisotropic design yield higher IPCE and optimized anisotropic design has an anisotropy index of one; implying that perfect anisotropy is favored. This design is characterized by narrow, wire-like PCBM clusters aligned in the direction of electrodes. These features contribute to a significant reduction in distance traversed by charges to respective electrodes and compensates the reduction in interfacial area. In contrast, the optimized isotropic design maximizes interfacial area, but its effect is subdued by tortuous paths for charge transport, leading to lower IPCE. Figure 3-7(B2 & C2) shows 150 x 150 x 150 voxel segments from the optimized anisotropic and isotropic microstructures and their corresponding SDFs. Anisotropic design, which is isotropic in XY plane but exhibits strong anisotropy along Z-axis, leads to an IPCE of 42.74% as compared to 41.58% of isotropic design. As suggested by Table 3-2, the difference in IPCE can be attributed to drastic reduction in distance traversed by electron to cathodes (S_C), a consequence of strong anisotropy of PCBM domains. The reduction in S_C dominates the minor increase in distance to nearest interface (d) observed in anisotropic design w.r.t isotropic design.

Table 3-2: Optimum design variables and resulting microstructural features

		Optimized Anisotropic Design	**Optimized Isotropic Design**
Optimum Design Variables	VF_{PCBM}	0.275	0.292
	k_i	1.45 nm^{-1}	1.80 nm^{-1}
	α	1	0
	$Average\ d$	0.28 nm	0.24 nm

Microstructural Features	$Average\ S_A$	$50.00\ nm$	$50.21\ nm$
	$Average\ S_C$	$50.00\ nm$	$60.82\ nm$

3.10 Summary

The new SDF based microstructure reconstruction methodology proposed is capable of reconstructing high resolution, isotropic/anisotropic microstructures to enable computational microstructure design in a wide range of materials systems. The representation is applicable for both isotropic and anisotropic microstructures with a small set of parameters that can serve as design variables in microstructure design. To quantify the level of anisotropy, we introduced anisotropy index, a scalar value bound in the interval [0,1] with 0 and 1 representing purely isotropic & anisotropic microstructures respectively. Meeting the need to reduce charge transport distances in the active layer, we formulated a design case study that pivots on SDF to find the optimal active layer microstructure with an aim to maximize IPCE. Only three design variables – PCBM volume fraction, anisotropy index and one SDF parameter were sufficient to represent microstructure, thus making optimization tractable. Optimization results reinforced our intuition that microstructures exhibiting strong anisotropy (anisotropic index ~ 1) in the direction of electrodes provide shorter paths for charges to travel towards respective electrodes; thus, delivering an enhancement in IPCE as compared to isotropic microstructures.

The dimensionality reduction and fast microstructure reconstruction afforded by SDF makes it a valuable tool in material design. In the next chapter, we shall once again leverage this method for the design of polymer nanocomposites microstructure but with an additional task of identifying the optimal composition concurrently.ng

4 Data Centric Design for concurrent composition and microstructure optimization

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 [80-82], modeling interphase behavior remains a prominent challenge. Researchers have investigated interphase behaviors and their origin both analytically and experimentally [82-84]. Recent experiments have demonstrated that the local polymer properties significantly change near the polymer surface via measurement of properties in model nanocomposites [84, 85]. 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 [80, 86-88]. 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) [8] techniques provide a quantitative representation of microstructures and the ability to reconstruct realizations with desired features. Among the existing methods, Physical Descriptors [21, 22] and Spectral Density Function (SDF) [10, 27, 46, 66] 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 [3, 4] - 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) [55, 89] has emerged as a viable proposition in material design [90-94]. 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) [95] 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 [96]. 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 [97-102] 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 [103] 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 [104-106].

In this chapter, we present 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.

4.1 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. Figure 4-1 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,

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.

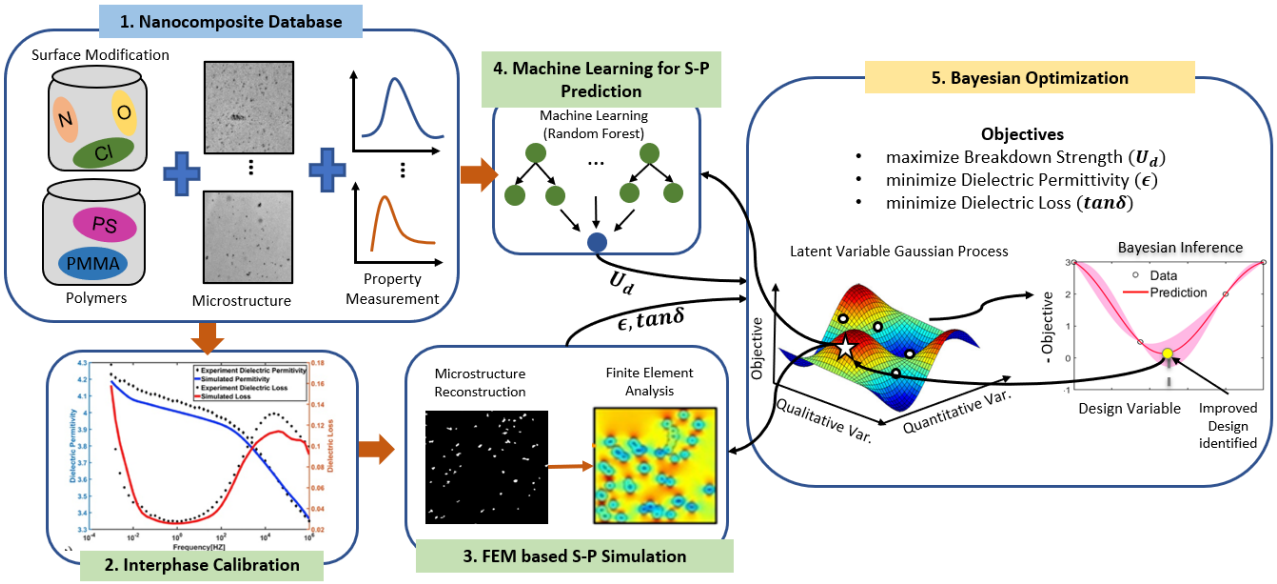


Figure 4-1: Data centric design framework for polymer nanocomposites

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 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. 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 [95, 96] 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.

We demonstrate the data-centric design process for electrically insulating polymer nanocomposites, with potential application in high voltage rotating machines [107]. 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 [108, 109].

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) [10, 27, 46, 66]. 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 [59] 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 [51] 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 [54] and expected maximin improvement [110] acquisition 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.

4.2 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.

4.2.1 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. [111]. The choice of polymer and surface modification determine nature of interactions between nanoparticle and polymer matrix. Our analysis [112] 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 [13, 113]. Dielectric spectroscopy

measurements was carried out for each nanocomposites sample prepared for this study, details of which is available in ref. [112, 114].

4.2.2 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. Although it is known to be the Fourier transform of a two-point autocorrelation function and hence encapsulates equivalent morphological information, Yu et al. [27] 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 [13] and assuming isotropy, SDF $\rho(k)$ was evaluated using Eq. (2-1). 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:

$$\rho(k) = \alpha * \exp\left(-\frac{k}{\theta}\right). \tag{4-1}$$

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. (4-1). 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. Figure 4-2 shows three microstructures along with their one-dimensional SDF and curve fitting. Filler dispersion increases through Figure

4-2(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.

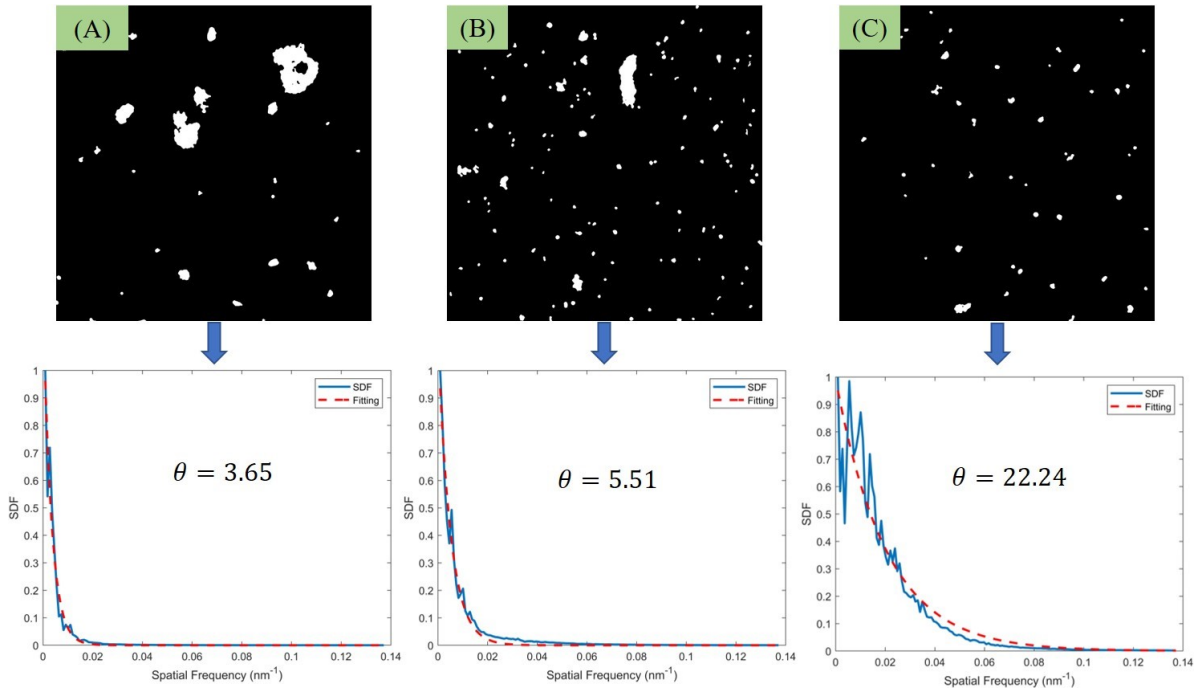


Figure 4-2: Three representative microstructures with varying dispersions and their SDF (blue curve) and corresponding curve fit using Eq. (4-1) (red dashed curve). The design variable θ's value for each image shown in inset.

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. [59] and described in Chapter 3.

4.2.3 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 [88]. 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 [80, 115]. Calibration of these scale factors is performed once for each of the six possible material combinations in Module 2. A detailed explanation of the calibration protocols is provided in Appendix 9.2.

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 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 4-1 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.

Table 4-1: Dielectric properties (relative to vacuum permittivity of 8.85×10^{-12} F/m) 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

4.2.4 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 [51] 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 (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 $R^2 = 0.92$ (Figure 4-3 (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 Figure 4-3(B).

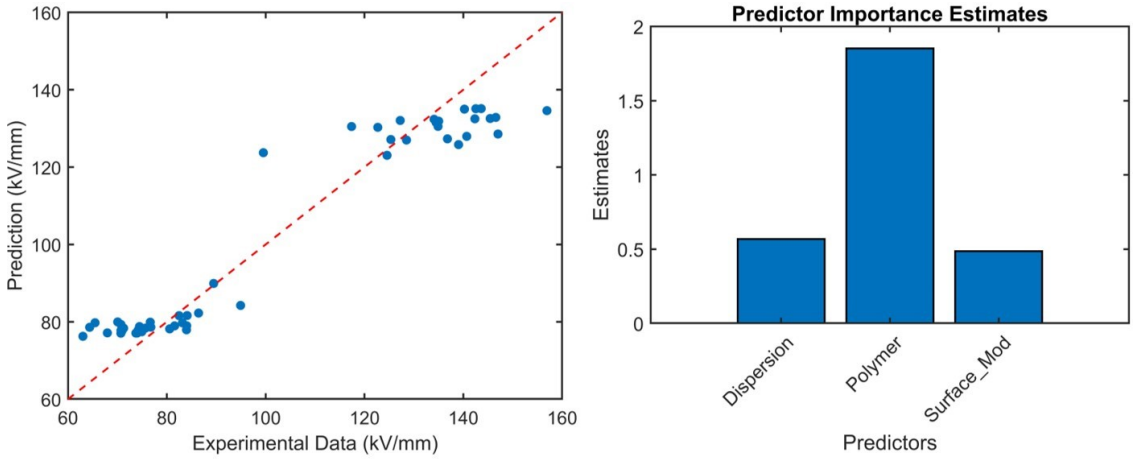


Figure 4-3: (A) Prediction accuracy of the random forest trained to predict breakdown strength. (B) Estimate of predictor importance deduced by random forest model. The larger the importance estimate for a predictor, the stronger its influence on breakdown strength.

4.2.5 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 the nanocomposite design problem under study which 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.

Consequently, we leverage LVGP as the machine learning model predicting the optimization objective(s) from the mixed design variables i.e., Step I of the BO procedure described in Chapter 2.4. We use LVGP models with two-dimensional latent space representation for all optimization results reported in Section 4.3. Uncertainty quantification provided by LVGP is used to accomplish Step II of the BO procedure as described below.

4.2.6 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 4-2. 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 4.2.2. We present both single and multicriteria BO strategies for this case study, using the same set of design variables with different objective formulations.

Table 4-2: 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]

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):

$$\begin{gathered}\min_{s\in S,p\in P,m\in M} \tan\delta^* + \epsilon^* - \mathrm{U}_\mathrm{d}^* \\ S: \{Chloro, Octyl, Amino\} \\ P: \{PMMA, PS\} \\ M: microstructures\ with\ 1.49 \le \theta \le 46.85,\end{gathered} \tag{4-2}$$

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 [54] is used as the acquisition function due to its ability to balance exploration and exploitation of design space, thus converging to optimum rapidly. Eq. (4-2) 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:

$$\min_{s\in S,p\in P,m\in M} \tan\delta^* + \epsilon^* - 10\mathrm{U}_\mathrm{d}^* \tag{4-3}$$

where S, P and M are the same as in Eq. (4-2). The modification of objective function subsequently affects the location of optimum in mixed-variable design space and will be discussed in Section 4.3.

Multicriteria optimization aims to find candidate designs lying on the Pareto frontier [116] – 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

$$\min_{\boldsymbol{w}\in W}\{y_1(\boldsymbol{w}), y_2(\boldsymbol{w}), \ldots, y_s(\boldsymbol{w})\}, \tag{4-4}$$

where $\boldsymbol{w}$ is the design input, W is the design space, s is the number of criteria, 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. (4-4) 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.(4-4), we use the expected maximin improvement (EMI) [110] acquisition function described as follows. Let the current Pareto set be composed of input set $P_W = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$ and output set $P_Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$, where k is the number of points in the Pareto set and $\boldsymbol{y}_i = [y_1(\boldsymbol{w}_i), y_2(\boldsymbol{w}_i), \ldots, y_s(\boldsymbol{w}_i)]^T, i = 1,2,\ldots,k$. For any given new input $\boldsymbol{w}_0$, the corresponding outputs are predicted by the LVGP models as $\boldsymbol{Y}_0(\boldsymbol{w}_0) = [Y_1(\boldsymbol{w}_0), Y_2(\boldsymbol{w}_0), \ldots, Y_s(\boldsymbol{w}_0)]^T$, where $Y_j(\boldsymbol{w}_0), j = 1,2,\ldots,s$ 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:

$$I\big(\boldsymbol{Y}_0(\boldsymbol{w}_0)\big) = \min_{\boldsymbol{w}_i\in P_W}\Big\{max\Big(\big\{y_j(\boldsymbol{w}_i) - Y_j(\boldsymbol{w}_0)\big\}_{j=1}^{s} \cup \{0\}\Big)\Big\}, \tag{4-5}$$

which is also a random variable. The larger the value of $I\left(\boldsymbol{Y}_0(\boldsymbol{w}_0)\right)$ 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 Figure 4-4, with one of the candidate points being $I(\boldsymbol{Y}_0) = 0$ and the other two points with a positive value $I(\boldsymbol{Y}_0)$.

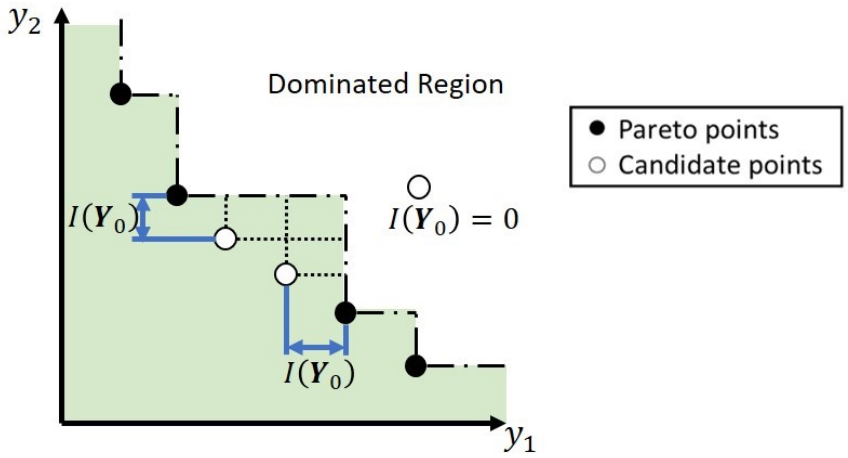


Figure 4-4: Values of the improvement metric $I(\boldsymbol{Y}_0)$ in a sampling process with two criteria.

The criterion for choosing the new evaluation input $\boldsymbol{w}_0^*$ is to maximize the expected value of improvement given in Eq. (4-5), i.e.,

$$\boldsymbol{w}_0^* = \underset{\boldsymbol{w}_0 \in W}{argmax} E(I(\boldsymbol{Y}_0(\boldsymbol{w}_0))). \tag{4-6}$$

When the original problem in Eq. (4-4) has mixed-variable input space W, Eq. (4-6) is a mixed-variable optimization problem. To solve Eq. (4-6), 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. (4-6), 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. (4-4) is adapted for multicriteria nanocomposite design as follows:

$$\begin{gathered} \min_{s\in S,p\in P,m\in M} \tan\delta, \epsilon, -U_d \\ S: \{Chloro, Octyl, Amino\} \\ P: \{PMMA, PS\} \\ M: microstructures\ with\ 1.49 \leq \theta \leq 46.85, \end{gathered} \tag{4-7}$$

where the variables have the same meaning as in Eq. (4-2). We use three independent LVGP models to predict the three dielectric properties from design variables S, P and M.

4.3 Optimization Results and Discussion

We performed 35 and 70 iterations of BO for single and multicriteria formulations respectively, as specified by Eq. (4-2) and Eq. (4-7) 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.

4.3.1 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 $U_d = 127.67 \frac{kV}{mm}$. Figure 4-5(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 polymer has a significant impact on the objective as indicated by Figure 4-5 (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) [117]. MATLAB's implementation of GA for mixed integer optimization was used in this study and applied to problem formulation defined by Eq. (4-2). For a fair comparison with BO, GA was configured to terminate after 65 objective function evaluations (seven generations with a population size of eight). Figure 4-5(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.

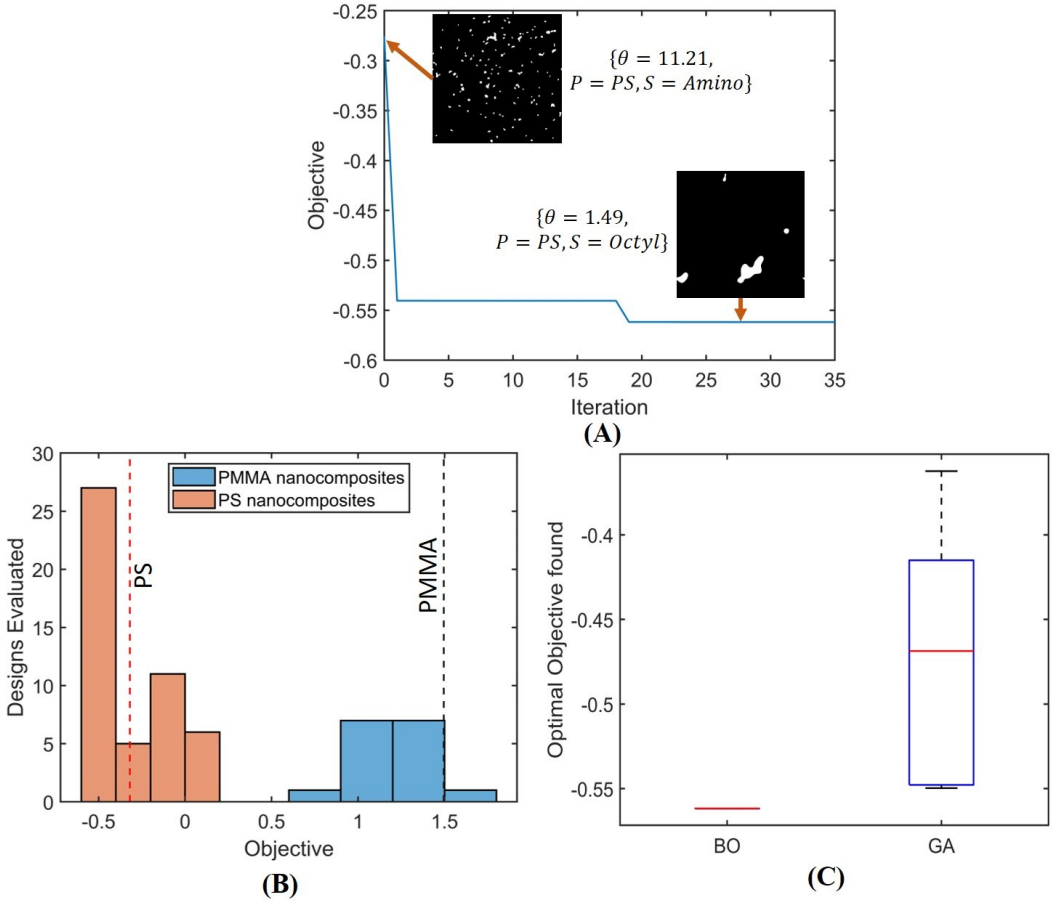


Figure 4-5: (A) Optimization history for single criterion BO that converged to objective = -0.562 along with three designs evaluated in the process (B) Distribution of evaluated designs, grouped by polymer type. Dashed lines denote objective values for PS & PMMA polymers (C) Comparison of ten replicates of BO and GA for single criterion optimization.

We also performed optimization using Eq. (4-3) 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. (4-2), 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.

4.3.2 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. Figure 4-6 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. (2-4).

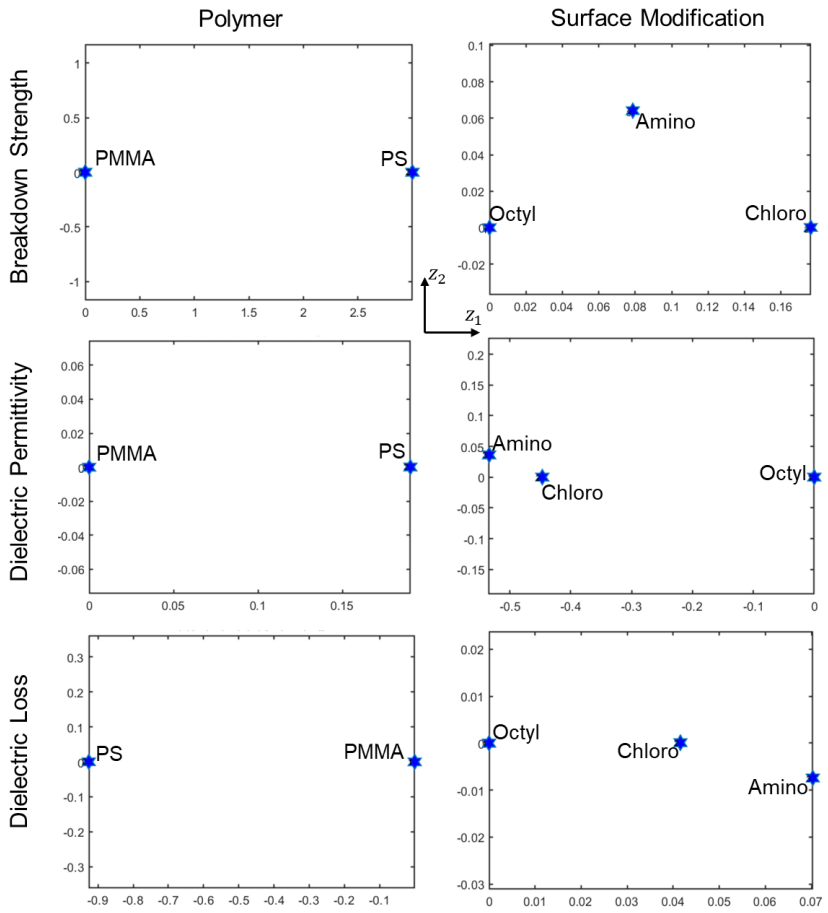


Figure 4-6: Visualization of latent variables for polymer and surface modification variables. Each row represents the latent variables estimated by the LVGP model used for corresponding property.

Figure 4-7 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 Figure 4-7) exclusively contains PMMA based samples and is not favorable to meet the design criteria. This is consistent with the findings in Figure 4-7 (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 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.

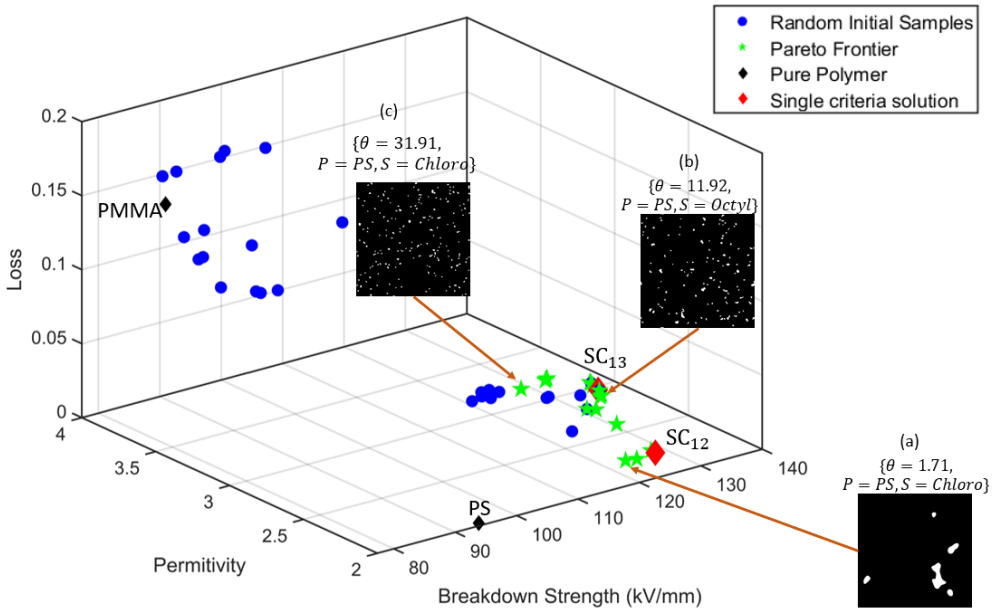


Figure 4-7: Summary of 70 iterations of Multicriteria Bayesian Optimization. SC_{12} and SC_{13} denote optimal single criterion solutions identified from Eq. (4-2) and Eq. (4-3) respectively.

The influence of design variables on dielectric properties is studied via Figure 4-8, 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 detailed analysis and testing depends on the modeler's preference based on the application, how the material is deployed, and device level performance.

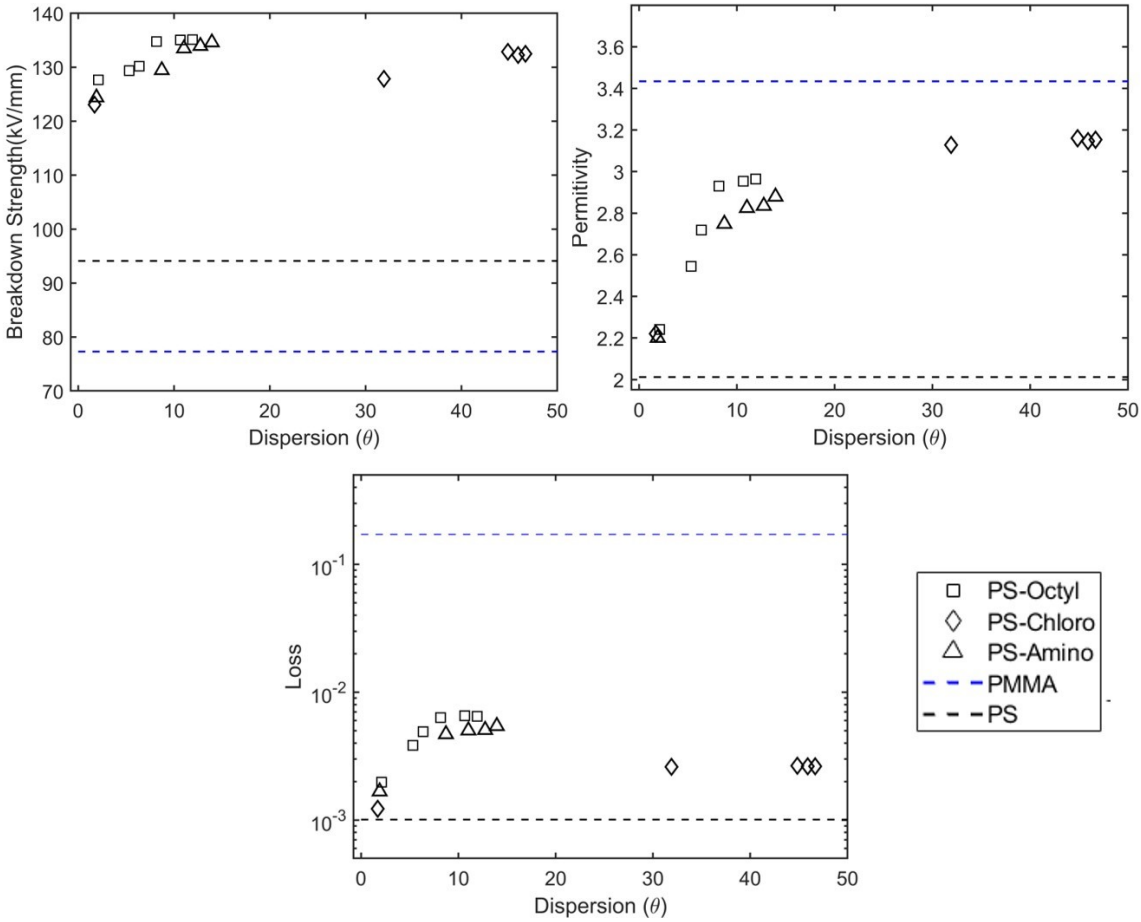


Figure 4-8: Influence of design variables on dielectric properties of nanocomposites on Pareto front. Dashed lines indicate property of polymer only system.

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 [113]:

$$\bar{I}_{\text{filler}} = f(\text{matrix}) \sinh^2(2\, W_{PF}/W_{FF} - 1)\log(E_\gamma + 1) + C_0, \qquad (4\text{-}8)$$

where $\bar{I}_{\text{filler}}$ is the normalized interphase area, $f(\text{matrix})$ and C_0 are polymer dependent constants, W_{PF}/W_{FF} is the filler-matrix compatibility descriptors and E_γ is the processing energy descriptor that we seek. For illustration, we choose the design (b) in Figure 4-7, 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, $f(\text{matrix})$ and C_0 are 0.00995 and 0.08798 respectively. For octyl-modified silica nanoparticles dispersed in PS, $W_{PF}/W_{FF} = 1.15$. Plugging these values in Eq. (4-8) 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.

4.4 Summary

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. This is further exemplified in the next chapter where we show that that BO with LVGP is ideally suited to tackle the combinatorial composition optimization problems arising in microelectronics design.

5 Featureless Combinatorial Optimization for Composition Design

In the previous chapter, we saw that concurrent design of composition and microstructure can be cast as a mixed variable optimization problem. LVGP allowed us to map the mixed variable inputs to the response and quantify uncertainty while BO used this information from LVGP to guide the search for promising designs efficiently. In some situations, the material properties may be entirely determined by the composition and thus materials design reduces to a combinatorial optimization problem which seeks to identify the best composition. Since the composition can be represented by a set of qualitative variables, we can once again leverage LVGP capability to map the qualitative inputs to the material properties and perform BO. In this chapter, we demonstrate one such example of composition optimization for discovery of novel metal-insulator-transition (MIT) compounds that are touted as promising alternatives to silicon for microelectronic devices. Through this example, we demonstrate a specific advantage that LVGP modelling provides for material design – a featureless machine learning approach that does not require domain expertise.

5.1 Introduction to Metal-Insulator-Transition compounds and design challenges

Upon traversing a critical temperature, the electrical resistivity of a MIT material can change by orders of magnitude [118]. Athermal approaches may also trigger the electronic transitions, including (chemical) pressure, variable carrier-densities, and applied electromagnetic fields. The transformations can be used to encode, store, and process information for beyond von-Neumann microelectronics and overcome performance limits of conventional field-effect transistors [119] for advanced logic/memory technologies [120]. Because macroscopic MITs occur in materials

with diverse chemistries and structures (Figure 5-1(a)[1]), various microscopic mechanisms – electron-lattice interactions, electron-electron interactions, or a combination thereof – lead to large variations in critical temperatures and accessible resistivity changes [121, 122]. This diversity exacerbates the efficient discovery and optimization challenge of achieving multiple property requirements to outperform silicon-based devices [123], including stability, large reversible resistivity changes (~10^5), and above room-temperature operation.

The aforementioned complexity is ubiquitous in formulating atomic scale material chemistry and macroscopic functionality relationships to guide property optimization. Presently, the principal solution relies on a better understanding of the underlying materials physics. Numerous data-driven machine learning models, however, have shown promising results in deciphering nonlinear relationships between materials structure and properties when sufficient training data is available [103, 124-128]. The predictive performance (error and efficiency) of these approaches is limited by the quality and quantity of the data, typically > $O(10^2)$, which poses a severe challenge to MIT materials design owing to the relatively small size of available dataset of ~ $O(10^1)$. The suitability of the machine learning model is determined by the input dimensionality and dataset size, which for high dimensional inputs necessitates large datasets and complex models for good predictive performance. A number of sequential materials design strategies have recently emerged [78, 128-130] to rescue the lack of data problem. Mostly being based on the Bayesian approach, these methods utilize knowledge extracted from existing data to infer properties of unknown materials following a step-by-step discovery manner. This sequential optimization method fits well

[1] All figures in this chapter are reproduced from Wang, Y., Iyer, A., Chen, W., & Rondinelli, J. M. (2020). Featureless adaptive optimization accelerates functional electronic materials design. *Applied Physics Reviews*, *7*(4), 041403, with the permission of AIP Publishing.

with the regular materials discovery procedure both experimentally and computationally, since property evaluations are usually time and effort consuming (e.g., synthesis and simulations). Nevertheless, these sequential learning models typically rely on numerical materials descriptors (features) whose selection may be informed by domain knowledge or trial-and-error approaches. For MIT materials systems which lack of microscopic understanding in how different compositions influence the phase transitions, this leads to ambiguity in feature formulation for discovery of MIT materials from structure and composition alone rather than through effective Hamiltonians [122].

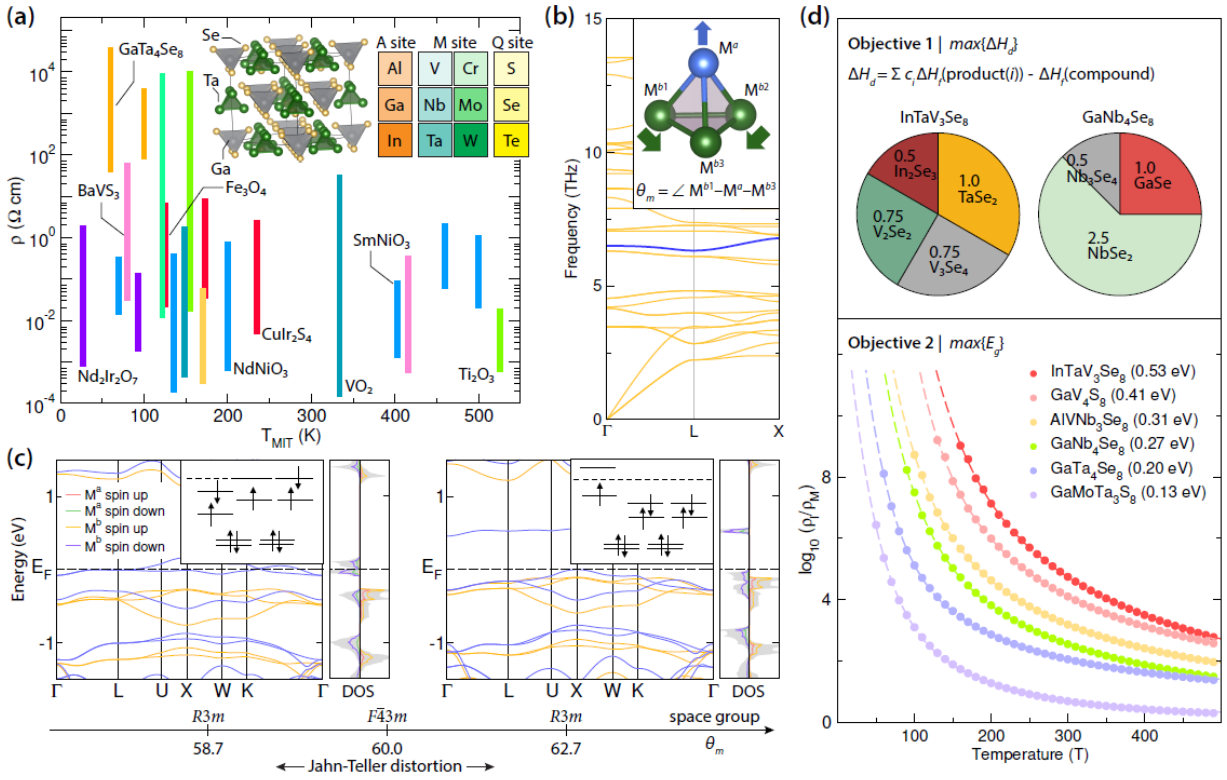


Figure 5-1: Metal-insulator transition materials and design objectives for the lacunar spinel family. (a) The range in resistivity accessible (length of bar) across the MIT and transition temperature for a variety of MIT materials. (left inset) The crystal structure of $GaTa_4Se_8$. (right inset) Candidate elements on each site of the lacunar spinel structure. (b) DFT-simulated phonon dispersion curves

of GaMo4S8 in the rhombohedral ground state, the blue curve corresponds to the Jahn-Teller active cluster distortion mode. (inset) The transition-metal cluster with a single apical M^a atom and three basal M^b atoms. The arrows indicate displacements characterizing the Jahn-Teller active phonon mode. The intra-tetrahedral cluster angle θ_m formed by $M^{b1} - M^a - M^{b2}$. (c) Electronic band structures and projected density of states (DOS in units of states/eV/spin/f.u.) of $GaMo_4S_8$ in its (right) semiconducting ground state and (left) metallic metastable phase with θ_m. The two $R3m$ phases are connected by the Jahn-Teller-type structural distortion with a $F\bar{4}3m$ intermediate state. (insets) Molecular orbital diagrams of the M_4 cluster with different local geometries. (d) Design Objective 1 is decomposition enthalpy change and the graphical decomposition pathways of two lacunar spinels is shown. The DFT-simulated temperature-dependent log ratio of the resistivity in the insulating and metallic phases of lacunar spinels serves as design Objective 2.

What could we do when there is little data available while the governing materials physics is not abundantly clear? Here we demonstrate a generic strategy to overcome the data scarcity as well as the feature engineering problems. We utilize multiobjective Bayesian optimization (MOBO) with latent-variable Gaussian processes (LVGP) to simultaneously optimize the band gap tunability and thermal stability in a family of candidate MIT materials – the lacunar spinels (introduced in the next section). With the goal to identify the optimal compositions, among hundreds of possible chemical combinatorics with both high functionality as well as synthesizability, we successfully retrieved all 12 superior compositions on the Pareto front by searching through a small fraction of the total design space. Notably, the chemical compositions (i.e., element on each crystallographic site) are all the model requires to guide this discovery procedure. No handcrafted features are required in this method, hence featureless learning, making

our methodology easily generalizable to other materials design problems. We also showcase how this model could offer helpful guidance on making better decisions towards the optimal design—selecting the next candidate compound to synthesize or simulate. Our adaptive optimization engine (AOE) frees researchers from exclusively relying on their chemical intuition, which can require an entire career to accumulate, and is particularly valuable when the research budget is limited.

5.2 Design Objectives

The complex lacunar spinel family $AM^aM_3{}^bQ_8$ with trivalent main group A, transition metal M, and chalcogenide Q ions demonstrate the complexity active in MIT materials design. The structure comprises transition-metal clusters (TMC) with M^a and M^b cations at the apical and basal positions of the tetrahedra (Figure 5-1(b) inset). Although there are hundreds of possible elemental combinations on the four lattice sites in the crystal structure (Figure 5-1(a)), only tens of the lacunar spinels have been experimentally reported [131, 132]. For example, GaV_4S_8 (M^a=M^b=V) exhibits a MIT [133], exotic spin textures [134], and multiferroism [135] while $GaVTi_3S_8$ shows negative magnetoresistance and half-metallic ferromagnetism [136]. Most lacunar spinels are narrow-bandwidth semiconductors in their ground states [131, 137]; these electronic properties are governed by distortions of the local TMC from the ideal T_d geometry [138], which manifest as low-frequency phonons as shown for $GaMo_4S_8$ (Figure 5-1(b), blue curve). Jahn-Teller-type distortions, which correspond to elongation along the [111] direction alter the TMC geometry, are particularly important; they transform the insulating $GaMo_4S_8$ ground state into a metastable metallic phase (Figure 5-1(c)). The MIT arises from a redistribution of electrons among the structure-driven orbital hierarchy (Figure 5-1(c) insets). Furthermore, these phases host low energy

electronic structures, discernible from the projected density of states (pDOS) in Figure 5-1(c), that arise from the different M^a and M^b sites. This capability to exhibit distinct and tunable electronic phases poses a challenge in the design of lacunar spinels from physics-based models while also making them an ideal system for MIT performance optimization.

In pursuit of novel MIT materials with superior performance, we specifically seek lacunar spinels that exhibit high thermodynamic stabilities and large resistivity-switching ratios, which we formulate as two design objectives for our materials discovery task. We reduce the approximately $O(10^3)$ compositional space to 270 candidates that maintain a $1M^a$ to $3M^b$ ratio. ($AM_2{}^aM_2{}^bQ_8$ compositions are excluded as they remove the C_3v symmetry fundamental to the MIT; Cr is also excluded from occupying the M^b site, because it destabilizes [139] the cluster.) This design space extends the known composition space that have been experimentally synthesized; therefore, it is important to determine the crystal stability, i.e., whether the selected chemical combination forms a thermodynamically stable lacunar spinel structure. To that end, we define the first design objective as the decomposition enthalpy change (ΔH_d, Figure 5-1(d)) and use density functional theory (DFT) simulations to evaluate formation energies (see Chapter 9.3). Materials with larger ΔH_d are expected to be more synthesizable [140] and stable during operation, making it a useful filter to prioritize compounds for subsequent theoretical analysis and synthetic processing. The second design objective is the ground state band gap (E_g). We use it as a proxy for the resistivity-switching ratio since E_g is positively correlated with the resistivity change between different electronic states (Figure 5-1(d)). A larger E_g also allows for greater band-gap tunability through control over the C_3v distortion, which is a desirable feature for programmable electronics. Importantly, because E_g is small for most MIT materials, stability is expected to be lower and more

difficult to achieve than that of nonpolymorphous compounds with majority ionic or covalent bonding [141].

5.3 Adaptive Optimization Engine

The nonlinear responses of both design objectives bring severe challenges to compound optimization beyond those amplified by chemical combinatorics using data-driven models. We overcome these obstacles by implementing a cyclic adaptive optimization engine shown in Figure 5-2, which consists of four iterative tasks (*vide infra*): property evaluation, aggregation of data (in a repository), featureless learning, and composition optimization. Beyond returning a predictive model capable of predicting properties from compositions alone, our iterative AOE leverages earlier approaches [78, 128, 129] to deliver materials with superior performance by design of composition-based solutions. In contrast to single objective design which often has a unique solution, multiobjective design aims to uncover the Pareto front—a set of non-dominated designs where no individual objective can be improved without deterioration in other objectives. In other words, the Pareto front represents the optimal trade-offs that can be achieved amongst competing objectives. There is no relative importance of multiple objectives in the process of identifying the Pareto front, which simply offers the designer several options from which to select the subset of compositions for further investigation and development. Since the designer's preference may be

subjective or informed by other criteria (e.g., cost), herein we present only the framework for Pareto front discovery and its comprising compositions.

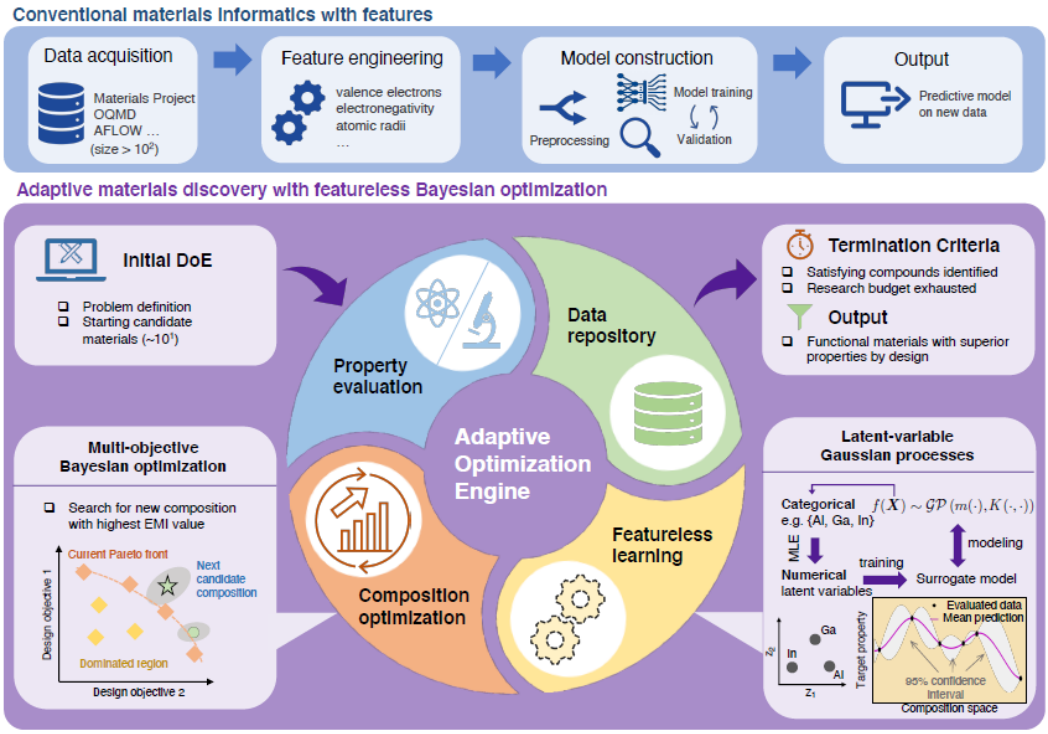


Figure 5-2: Comparison of conventional (feature-required) machine learning with the featureless adaptive optimization engine. Upper panel: The workflow of a conventional feature-based machine learning model typically involves data acquisition, feature engineering, model construction, and property prediction. Lower panel: The adaptive materials discovery scheme.

The AOE has the important advantage of bypassing the feature engineering procedure as in conventional ML methods; it learns properties directly from the chemical composition at each site (i.e., A, M^a, M^b, Q). Gaussian Process (GP) is ideally suited for this problem, because (a) it interpolates data and hence is ideal for surrogating deterministic responses such as DFT results, and (b) it provides a principled statistical representation for uncertainty quantification, which is essential for Bayesian optimization. Latent-variable methods provide a fundamentally different

approach to modelling categorical design variables by alleviating the need for handcrafted features. It transforms categorical variables (i.e., elemental compositions) into a continuous numerical space. Utilizing these approaches in the AOE, we achieve featureless learning and then perform composition optimization under the multiple objectives through LVGP.

We start the MIT-materials AOE for the lacunar spinel family through an initial design of experiment (DoE) consisting of four experimentally known compounds within the family (i.e., $GaMo_4S_8$, GaV_4S_8, $GaNb_4Se_8$, and $GaTa_4Se_8$) and eight new compositions generated by discretized Latin Hypercube Design (LHD) [142] (Figure 5-3). A four-dimensional Latin Hypercube Design of size eight is generated, where each dimension corresponds to a crystal site (e.g. A, M^a, etc.). Since the four known compounds are all gallium-based, we only consider Al and In for the A site design. (b, c) Each dimension is evenly divided into a number of grids, each grid represents one candidate elemental composition at that crystal site. For instance, the Q site is divided into three grids because there are three candidate elements (S, Se, Te) on that site. The designed composition could then be determined using the grid-composition correspondence. For example, Design ID Number 1 (D1) resides in the grid corresponding to {Al, Mo, V, S}; therefore, its composition is $AlMoV_3S_8$. This procedure ensures a variety of elemental combinations within the initial DoE set, where each candidate element will appear at least once, so that the model has knowledge about different elemental contributions to the design objectives.

(a)

Design	A site	M^a site	M^b site	Q site
D1	0.281 Al	0.241 Mo	0.595 V	0.017 S
D2	0.074 Al	0.253 Mo	0.682 Ta	0.665 Se
D3	0.718 In	0.802 W	0.862 W	0.875 Te
D4	0.183 Al	0.485 V	0.140 Nb	0.560 Se
D5	0.486 Al	0.007 Nb	0.297 Mo	0.791 Te
D6	0.947 In	0.510 Ta	0.446 V	0.418 Se
D7	0.511 In	0.635 Ta	0.972 W	0.270 S
D8	0.787 In	0.934 Cr	0.116 Nb	0.182 S

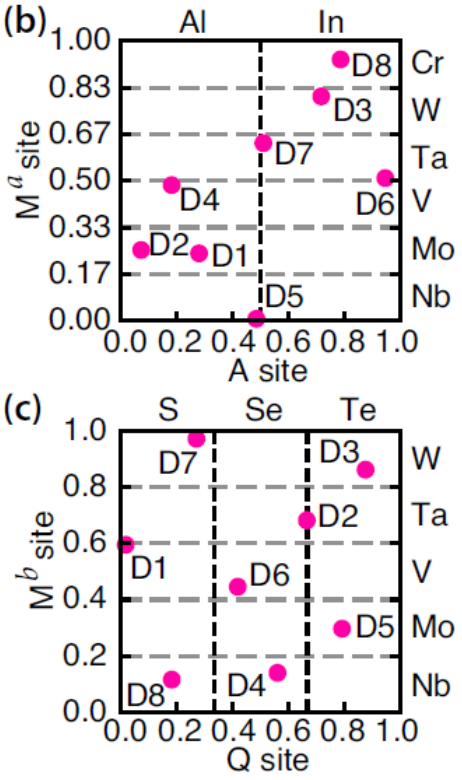


Figure 5-3: Design of experiment (DoE) for the complex lacunar spinel family. (a) A four dimensional LHD of size eight and its mapping to crystal sites for the lacunar spinel problem. (b,c) Location of designs listed in (a) in the four dimensional space which is discretized based on number of levels to be allocated along each dimension.

Next, we use high-fidelity DFT simulations to evaluate ΔH_d and E_g (see Chapter 9.3). This is the most resource-intensive step among the four tasks; therefore, it is desirable to iterate through the AOE (property evaluation) step as few times as possible. Although it is application dependent, AOE can be terminated if a compound with target properties is discovered, or the budget (computational/experimental) has been exhausted. Then, we create a data repository that contains entries for both composition and the evaluated properties. Unlike other ML methods, we do not rely on a large number of existing data at either the onset or later in the learning process.

We then construct a LVGP model by mapping the elemental compositions (e.g., Al, Ga, In) into a two-dimensional (2D) latent space (Figure 5-2 lower right inset) where the relative positions of elements are obtained using maximum likelihood estimation (MLE). This latent space representation enables us to construct Gaussian process surrogate models for the unknown underlying design objectives, ΔH_d and E_g, as a function of composition. The MOBO step then begins, and we use the LVGP models to predict ΔH_d and E_g of the unexplored compositions in our design space; we choose the next candidate composition for evaluation using the expected maximin improvement (EMI) as the acquisition function, which quantitatively describes the performance gain compared against the compositions at the current Pareto front. The EMI is defined in such a way that both objectives have equal weighting, and the objective properties are normalized with respect to the current min-max values. This acquisition function considers both exploration of compositions with high uncertainty (Figure 5-2, shaded ellipses, lower left inset) as well as exploitation of candidates with high performance gain. The composition with highest EMI is then selected for DFT simulation (property evaluation), at which point another AOE cycle commences.

The aforementioned iterative optimization procedure progresses and explores the available design space. One new lacunar spinel composition is evaluated and added to repository after each AOE iteration. The LVGP models are also updated in each iteration as more knowledge becomes available. Owing to the high computational cost of the property evaluation process, we terminate the optimization process after searching through 1/3 of the entire design space. In order to validate the effectiveness of this method, we ultimately evaluated ΔH_d and E_g with DFT calculations of all 270 compositions within the design space by expending approximately 3×10^6 CPU hours.

5.3.1 Adaptive Optimization Engine Implementation

The adaptive materials discovery scheme starts from an initial set of design of experiments (DoE), where system variables, design objectives, and design space are first defined for the problem, providing a few $O(10^1)$ candidate materials to initialize the discovery procedure. The target material properties (design objectives) are evaluated either by experimental measurement or theoretical simulations. Candidate composition and its evaluated properties are then added to a data repository, which initially may either be empty or only contains entries for existing materials within the design space. Its size grows as more candidate materials are evaluated during the adaptive optimization process. Featureless learning involves directly learning from the chemical composition of materials comprising the data repository by mapping each compositional variable into a two-dimensional latent space (spanned by z_1 and z_2) using maximum likelihood estimation, which enables the construction of a LVGP surrogate model. Two independent LVGP models with Gaussian correlation function are fit at each iteration to predict E_g and ΔH_d, respectively. In each LVGP model, categorical variables A, M^a, M^b and Q are represented by a 2D numerical latent variable vector to evaluate their correlation. Note that each categorical variable resides in its unique latent space. For the LVGP model predicting E_g, let $\boldsymbol{z}^A = [z_1^A, z_2^A]$ denote the latent variable for the A site. Similarly, $\boldsymbol{z}^{Ma}$, $\boldsymbol{z}^{Mb}$ and $\boldsymbol{z}^Q$ denote the latent variables for M^a, M^b and Q site, respectively. Then, the Gaussian correlation (r) between E_g of two compounds, e.g., $GaMoV_3S_8$ and $AlNbW_3Se_8$, is:

$$r\left(E_g^{GaMoV_3S_8}, E_g^{AlNbW_3S_8}\right) = \exp\left(-\left|\left|z^{Ga} - z^{Al}\right|\right|_2^2 - \left|\left|z^{Mo} - z^{Nb}\right|\right|_2^2 - \left|\left|z^{V} - z^{W}\right|\right|_2^2 - \left|\left|z^{S} - z^{Se}\right|\right|_2^2\right), \tag{5-1}$$

where $||.||_2^2$ represents the Euclidean 2-norm. This procedure is used to compute the correlation matrix for properties of all evaluated compositions. The positioning of latent variables $\boldsymbol{z}^A$, $\boldsymbol{z}^{M_a}$, $\boldsymbol{z}^{M_b}$, and $\boldsymbol{z}^Q$ in their corresponding latent space are estimated via MLE as described in Chapter 2.3. The LVGP model for ΔH_d also utilizes the 2D latent variable representation $\boldsymbol{\kappa}^A$, $\boldsymbol{\kappa}^{M_a}$, $\boldsymbol{\kappa}^{M_b}$, and $\boldsymbol{\kappa}^Q$ as previously defined to evaluate the correlation $r\left(\Delta H_d^{GaMoV_3S_8}, \Delta H_d^{AlNbW_3Se_8}\right)$ in a similar manner. Multi-objective Bayesian optimization is then performed with the LVGP models to obtain the next candidate material composition with the highest expected maximin improvement (EMI) value. Multiobjective Bayesian optimization includes first defining the lacunar spinel family $A\mathrm{M}^a\mathrm{M}_3{}^b Q_8$ with $A \in$ {Al, Ga, In}, $\mathrm{M}^a \in$ {V, Nb, Ta, Cr, Mo, W}, $\mathrm{M}^b \in$ {V, Nb, Ta, Mo, W} and $Q \in$ {S, Se, Te}. The design space ($\boldsymbol{C}$) comprises 270 compounds, each compound is represented by four design variables A, M^a, M^b and Q with three, six, five, and three choices, respectively. Our objective is to maximize E_g and ΔH_d, which is represented in standard optimization formulation as:

$$\min_{\boldsymbol{c}\in\boldsymbol{C}} -E_g(\boldsymbol{c}), -\Delta \mathrm{H_d}(\boldsymbol{c}), \tag{5-2}$$

Starting from the initial dataset, the AOE evaluates new candidate compounds by gauging their improvement in the design objectives. Here, we use the expected maximin improvement (EMI)

metric [110] to guide the adaptive sampling framework. The Maximin Improvement (I_M) for compound $\boldsymbol{c}$ is:

$$I_M(\boldsymbol{c}) = \min_{\boldsymbol{c_i} \in \boldsymbol{C_{PF}}} \left\{ \max\left(\widetilde{E_g}(\boldsymbol{c}) - \widetilde{E_g}(\boldsymbol{c_i}), \Delta\widetilde{H}_d(\boldsymbol{c}) - \Delta\widetilde{H}_d(\boldsymbol{c_i}), 0 \right) \right\}, \tag{5-3}$$

where $\boldsymbol{C}_{PF}$ is the current set of Pareto front compositions. To facilitate the comparison of objectives in Eq. (5-3), we scale the value of each design objective P using the scheme $P(\cdot) = \frac{P(\cdot) - P_{min}}{P_{max} - P_{min}}$ where $P_{\max}$ and $P_{\min}$ are the maximum and minimum value of property observed so far. By scaling the properties, we ensure all design objectives are comparable and viewed equally. The EMI of compound $\boldsymbol{c}$ is defined as the expected value of I_M:

$$\mathrm{EMI}(\boldsymbol{c}) = \mathbb{E}[I_M(\boldsymbol{c})], \tag{5-4}$$

We evaluate the EMI through Monte Carlo sampling with 500 trials. At each AOE iteration, the EMI is calculated for all compositions that are not yet present in the data repository. The composition with largest EMI will be sampled next in property evaluation and then added to the data repository. Figure 5-2 highlights these steps. The model accounts for uncertainty with the 95% confidence interval shown as the shadowed area around the new compositions (the green symbols). In the lower left inset, the green star composition outperforms the green circle composition, and will be passed to the next property evaluation procedure. The iterative optimization step continues until all compounds satisfying the objectives are discovered, forming the Pareto front, or computational resources expire.

5.3.2 Adaptive Optimization Engine Performance

Figure 5-4(a) displays the results of the AOE. We successfully identify all 12 materials at the true Pareto front within 53 iterations (red asterisks, upper panel)—compositions and objective-

related properties are enumerated in Figure 5-4. Combined with the 12 compounds from our initial DoE, we explored less than 25% of the entire design space before identifying all lacunar spinels on the Pareto front. Interestingly, Pareto-front compositions are mostly found with high EMI values, showing that our model makes beneficial recommendations on which composition to evaluate next. High prediction uncertainty likely explains why a Pareto-front composition is not identified for some iterations with a large EMI. The EMI values reduce to nearly zero after all Pareto front compositions are identified (blue, upper panel) since all candidates not sampled are dominated by the Pareto front compounds. We also show the absolute error in the LVGP-predicted ΔH_d (pink) and E_g (orange) values of the evaluated composition at each iteration to further demonstrate the effectiveness of our model (Figure 5-4 (a)). We find a general decreasing trend in error and therefore better model predictability as it becomes aware of more composition-property knowledge.

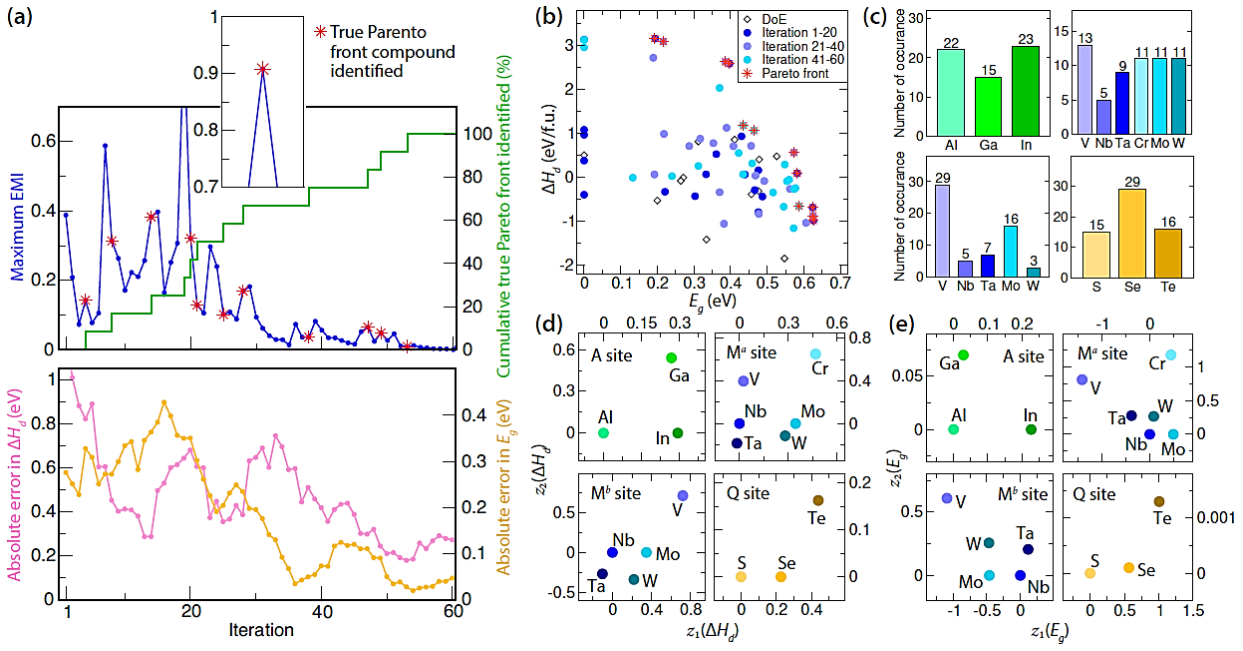


Figure 5-4:The results of adaptive optimization on the lacunar spinel family**.** (a), Upper panel: Evolution of the highest expected maximin improvement (EMI, blue line) and percentage of true

Pareto front compounds identified (green line) as a function of iteration number. Results of the first 60 iterations are shown here. The red asterisks represent sampling points where a true Pareto front design is successfully identified. Lower panel: The moving average of absolute error in the predicted E_g and ΔH_d values for a compound selected by the acquisition function for property evaluation. (b), The distribution of initial design of experiment and the first 60 evaluated compounds. Compounds evaluated in earlier stages have darker colors. True Pareto front designs are marked with red stars. (c), Distribution of Bayesian optimization-sampled elemental compositions for the first 60 iterations. (d, e), Latent space representation of elemental composition at different crystal structure sites in the ΔH_d and E_g surrogate model, respectively. Results obtained after 60 iterations.

Figure 5-4 (b) shows the history of composition explored by the AOE for the first 60 iterations. The initial DoE sets are relatively scarcely distributed away from the true Pareto front (marked as red asterisks), yet the model explores regions far from that covered by the DoE sets and is able to identify 75% of Pareto front compositions within the first 40 iterations. First, we begin to understand this performance by examining the distribution of elements sampled by the MOBO (Figure 5-4 (c)). Our model does not exhibit much compositional bias upon sampling elements for the A site; however, it shows clear preferences for choosing certain elements on other sites. V and Mo are sampled more frequently on the basal M^b site, while Nb and Ta are less favored on the apical M^a site. Se is also preferred over S and Te for the Q site.

Then we examine the 2D latent space representations for both design objectives obtained after 60 iterations of AOE (Figure 5-4 (d) and (e)). The relative positioning of elements in the latent space reflects correlations in their influence on properties; elements in close proximity exhibit

similar impact. Interestingly, different transition metals exhibit distinct correlation patterns across various sites and objective properties. This variation leads us to conclude that (i) the transition metals contribute to stability and band gap in different and unexpected ways, and (ii) the lack of any resemblance in element positioning in the site-dependent latent spaces, except for the M^a site, to the periodic table indicates that chemical-intuition-based MIT design within the lacunar spinels is highly nontrivial. For example, chromium is located far from the other elements in the M^a latent space, indicating that its influence on properties is distinct. Indeed, Cr containing compounds have significantly lower E_g and higher ΔH_d (Figure 5-5).

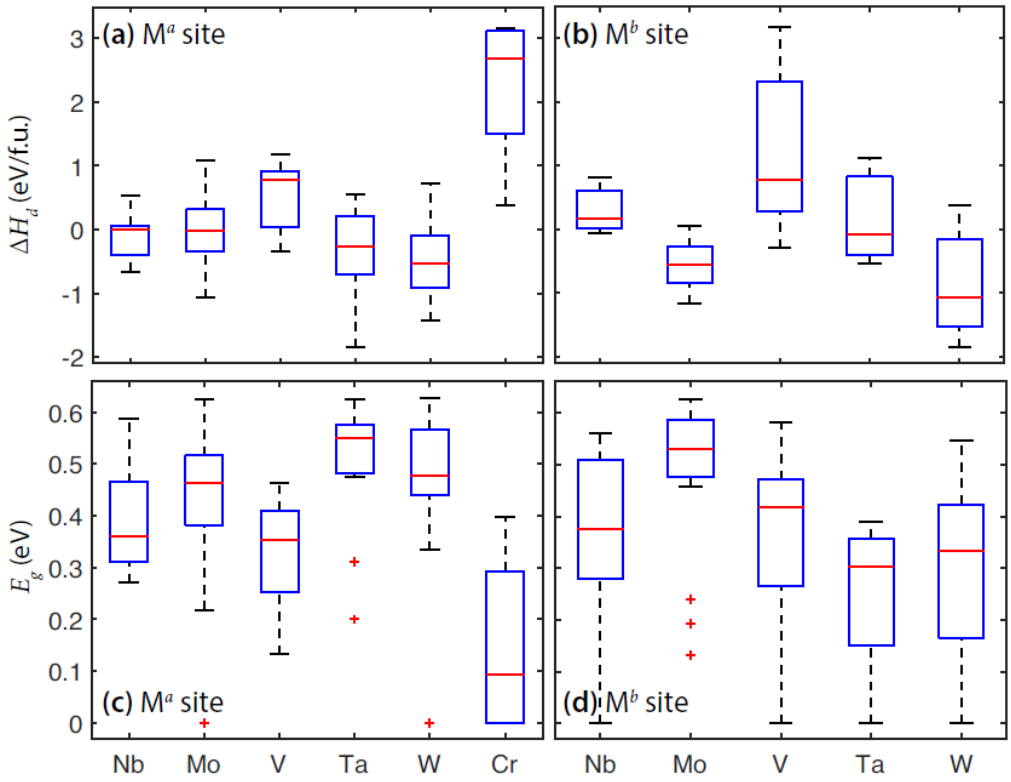


Figure 5-5: Composition-property relationships at the transition-metal sites**.** Distribution of DFT-evaluated properties of the complex lacunar spinel family with 12 initial DoE sets and 60 iterations of AOE. This data presents the impact different elemental compositions at the transition metal sites

(i.e., M^a and M^b) have on the two design objectives (i.e., ΔH_d *and* E_g). (a, b) decomposition enthalpy change distribution at M^a, M^b site. (c, d) band gap distribution at M^a, M^b site.

The aforementioned performance is robust as revealed by our multi-trial results (Figure 5-6(a)), where we find the AOE successfully identifies 90% of the true Pareto-front compositions by exploring 30% of the design space with different initial DoE sets. In each trial, the initial DoE set consists of the same four known lacunar spinel compounds and eight new compositions designed by the DoE procedure. Since LHD is inherently random, repeating the DoE procedure will lead to another randomly generated DoE set. Therefore, we use this method to run multiple trials of AOE with different DoE sets. The size of DoE is another parameter for the designer to select in the AOE framework. Since the computational budget is often the bottleneck in discovery, the designer must allocate it wisely between the DoE and AOE. We investigated this problem using a set of four DoE sizes: 6, 12, 18, and 24, because there are six elements admissible at the M^a site (Figure 5-6 (b)). In each case, the computational budget is fixed to 40 and 60 simulations, and they are split between DoE size and AOE iterations. For example, 40 simulations can be split into DoE of size 6 and 34 iterations of AOE whereas a DoE of size 12 corresponds to 28 iterations of AOE, etc. Here, the four known gallium based compounds were not explicitly included in the DoE. We find that using a small DoE to initialize AOE (conversely, allocating more simulations to the AOE) is advisable, as its uncertainty guided exploration is more likely to discover Pareto compositions (Figure 5-6 (b)).

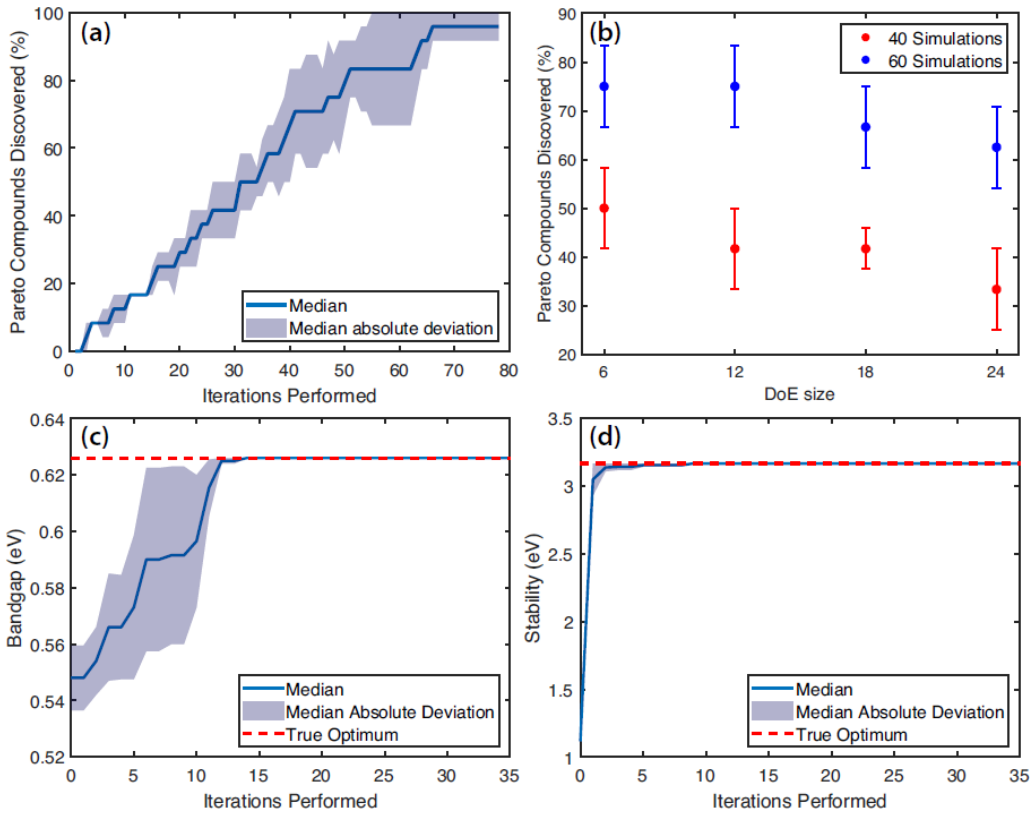


Figure 5-6: Robustness of the Adaptive Optimization Engine (AOE). (a) The optimization history for 10 replicates of AOE, each initialized with a distinct set of 12 initial DoE compounds. Solid line shows the median percentage of true Pareto front compounds discovered at each iteration. The shaded area represents the median absolute deviation across 10 trials. (b) The fraction of Pareto front compounds discovered when the computational budget is fixed to 40 and 60 simulations. Filled circles and their corresponding error bars represent the median and median absolute deviation respectively. (c,d) The optimization history of 10 replicates of single-objective Bayesian optimization, targeting maximum band gap (E_g) and stability (ΔH_d), respectively. The initialization method is the same as described in (a). Global optimum ($E_g^* = 0.626\ eV, \Delta H_d^* = 3.167\ eV$) is identified within 10 % exploration of design space.

Single-objective Bayesian optimization on both band gap (E_g) and stability (ΔH_d) are also performed using Expected Improvement acquisition criterion [54], as shown in Figure 5-6 (b, c), respectively. Unsurprisingly, the model shows much higher efficiency in identifying the optimal composition than in the multi-objective task, where less than 10% of the entire design space is explored. We also notice that the model is always able to quickly infer the compound with highest stability, as depicted by the steep curve in Figure 5-6 (c). Intuitively, thermodynamic stability is straightforward to linearize from elemental reference states whereas the band gap is determined by the valence electronic structure and multiple interactions. Therefore, it might be easier for the model to decode the relationship between composition and stability, while learning the band gap dependency requires accumulating more knowledge.

5.3.3 Pareto Compound Analysis

We use DFT simulations to examine the properties of the identified Pareto-front compositions, focusing on $\Delta H_d, E_g$ and the Jahn-Teller active phonon ν_{JT} involved in the MIT (Table 5-1). We find most Pareto-front compositions consist of two different cations on the M^a and M^b site, only three have $M^a = M^b$, with 75% of the optimized materials being selenides. GaV_4Se_8 is the only Pareto front compound previously synthesized, and verified to exhibit resistive-switching behavior under an applied electric pulse [143]. All compounds exhibit $R3m$ symmetry and are dynamically stable in their ground state ($\nu_{JT} > 0$). The phonon frequencies of the selenides, including ν_{JT} are lower than those of the sulfides. All of the designed lacunar spinels also exhibit semiconducting gaps with semilocal exchange correlation and static Coulomb interactions and exhibit nonzero electric polarizations. Compositions with larger band gaps tend to have lower stability as determined by ΔH_d: 2/3 are stable ($\Delta H_d > 0$, indicating decomposition is endothermic), whereas

four of the 12 compounds comprising Mo have small values of $\Delta H_d < 0$, which could nonetheless be stable and synthesizable [140, 144]. Typically, highly ionic materials with large electronic band gaps are also quite stable (e.g., NaCl). However, we find a clear trade-off between these two properties for the Pareto front compositions. One possible reason is because all of these candidate materials are small-gap semiconductors (with $E_g < 0.65eV$) due to metal-metal and semiconvalent bonding while also being polymorphous; therefore, these lacunar spinels are unlikely to follow the general trend. In addition, Figure 5-7 shows that the transition metals contribute to E_g and ΔH_d in quite different ways, which could lead to this functionality-stability trade-off. The AOE, however, does not possess knowledge of chemistry beyond the lacunar spinel family; yet it is able to resolve the $\Delta H_d - E_g$ relationship regardless of whether there is a trade-off or positive correlation. These findings reinforce the effectiveness of this model. Although the ground states of these materials are all semiconducting, we find two different electronic transitions upon traversing the ideal TMC geometry ($\theta_m = 60°$): the expected (Type I) metal-to-insulator transition and an unexpected (Type II) semiconductor-to-insulator transition (SIT). Figure 5-7(a) shows the changes to the electronic structure for the MIT lacunar spinels $AlTaV_3Se_8$ and $InWMo_3Se_8$ with the insulating state (lower panel) always lower in energy than the metastable metallic phase (upper panel) after the Jahn-Teller-type distortion ($\theta_m \neq 60°$, Figure 5-7). The pDOS of these compounds show that the metallic state in the Type I transition arises from cluster distortion-triggered orbital ordering and occupancy changes, similar to the mechanism depicted in Figure 5-7(b). However, the metallic states are different owing to the chemistry of the metals comprising the TMCs. We also find that the basal M^b site plays a more decisive role near the Fermi level with minor contribution from the apical M^a site. The M^a site on the other hand, plays an

active role in the Jahn-Teller-active phonon owing to differences in atomic mass (Figure 5-7). The remaining lacunar spinels in Figure 5-7(a), $InNbMo_3Se_8$, $InTaMo_3Se_8$, $InCrV_3S_8$, and $InWV_3S_8$, exhibit a Type II transition. The lower and upper panel show their ground and metastable state pDOS, respectively. Interestingly, some compounds undergo singlet formation and transform into a nonmagnetic phase (e.g., $InNbMo_3Se_8$) while others remain ferromagnetic after the cluster distortion (e.g., $InCrV_3S_8$) owing to competition between spin-pairing and magnetic interactions [145].

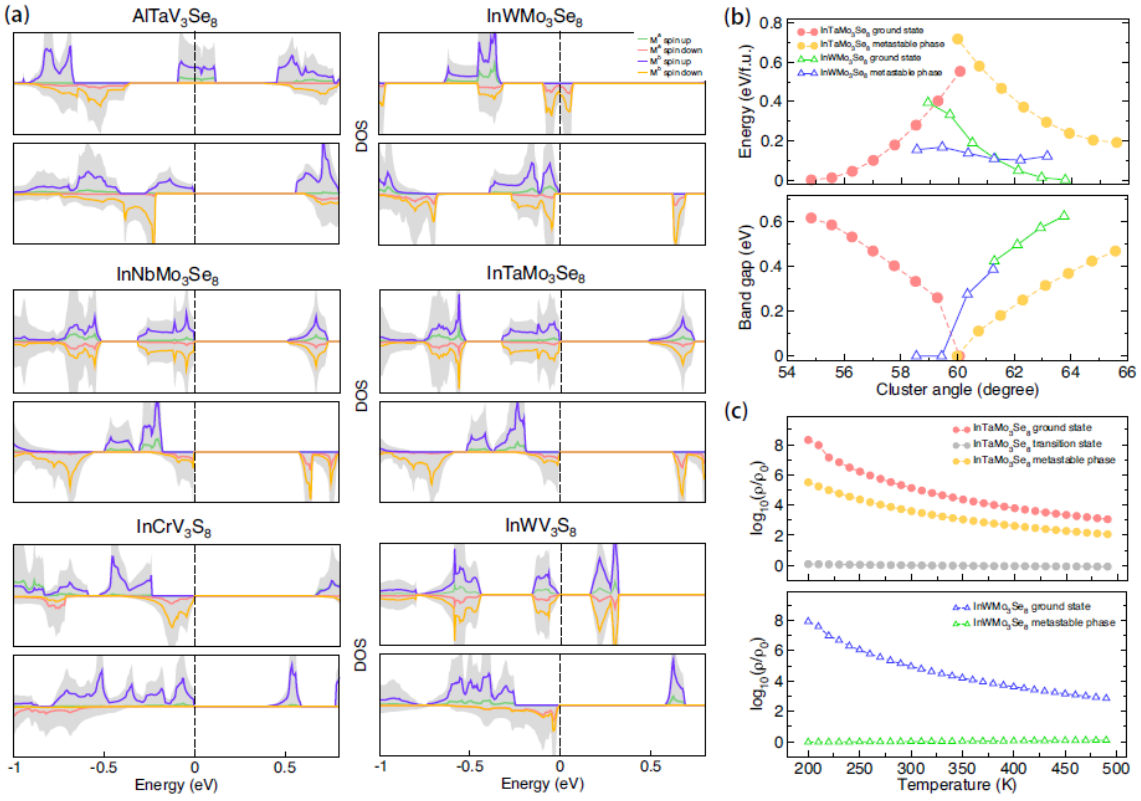


Figure 5-7: DFT-simulated electronic properties of selected lacunar spinel compositions at the Pareto front**.** (a) The projected electronic density of-states (DOS) of $AlTaV_3Se_8$, $InWMo_3Se_8$, $InNbMo_3Se_8$, $InTaMo_3Se_8$, $InCrV_3S_8$, and $InWV_3S_8$. The lower panel of each composition shows the ground state electronic structure and the upper panel shows the DOS of the metastable phase

after the Jahn-Teller distortion. Both panels are normalized and span a range of 15 states per formula unit for each spin channel (vertical axis). $AlTaV_3Se_8$, $InWMo_3Se_8$ exhibit metal-insulator transitions whereas the other compounds show semiconductor-to-insulator transitions. (b) The DFT relative energies and band gaps of $InWMo_3Se_8$ and $InTaMo_3Se_8$ as a function of the cluster distortion angle θ_m. $InTaMo_3Se_8$ undergoes a semiconductor-to-insulator transition with a metallic intermediate state for $\theta_m = 60°$. (c) Simulated DC resistivity of the compounds in (b) for their corresponding metallic, semiconducting, and intermediate states.

Last, we model the switching process and resistivity upon structural distortion for $InWMo_3Se_8$ (Type I) and $InTaMo_3Se_8$ (Type II) by modulating the amplitude of the ν_{JT} atomic displacements for each material in both the (insulating) ground and (metallic or semiconducting) metastable states. The DFT simulated energy and corresponding band gap at different cluster angles (θ_m) are shown in Figure 5-7 (b). Both compounds show first-order transitions. Owing to the small changes in the TMC geometry required for switching, readily available external stimuli could be used to trigger the transitions [133, 146, 147]. The simulated DC resistivity of $InWMo_3Se_8$ and $InTaMo_3Se_8$ clearly shows the promising functionality of these newly discovered compositions in the lacunar spinel family (Figure 5-7(c)). Since we successfully identify all 12 Pareto-front compositions by searching through less than 25% of the design space, our work demonstrates the efficiency of featureless adaptive materials discovery for electronic materials design. The featureless AOE is particularly useful when data availability and physical understanding of the target materials system is limited at either the atomic or microstructural scale.

Our multiple property objectives of high stability and large insulating band gaps were achieved by using Bayesian optimization (BO) for MIT materials-composition design without explicitly

constructing features (descriptors) via latent-variable Gaussian process implemented in our adaptive optimization engine. We successfully identified all 12 Pareto-front lacunar spinel compositions by searching through less than 25% of the design space. Since the Utopian composition with both high functionality and stability (i.e., the upper right corner of Figure 5-4(b)) cannot be realized, the Pareto front illustrates the trade-offs among objectives. This information is beneficial to materials scientist as it aids in the selection of candidate materials to further investigate or deploy. The selection rules will depend on the designer's preferences and whether to favor one property over others as well as their willingness to compromise. Therefore, we report the steps needed to identify all Pareto designs to quantify our model efficiency. Because these materials have garnered much research attention in recent years owing to the richness of their fascinating physical behaviors (e.g., MITs, skyrmion lattices, and superconductivity), we anticipate the newly identified lacunar spinels will be pursued experimentally in search of these phenomena.

Table 5-1: DFT-evaluated ground state properties of the Pareto front compounds. NOI is the number of iterations taken to discover the compound during the adaptive optimization process. Values of $\Delta H_d > 0$ (units of eV f.u.$^{-1}$) indicate an endothermic reaction occurs and the stable compound disfavors decomposition. E_g is the DFT band gap in eV. ν_{JT} is the frequency (THz) of the Jahn-Teller-type phonon involving the TMC. P is the electric polarization in μC cm^{-2}. The value of θ_m in the insulating ground state and transition type, Type I (MIT) or Type II (SIT), are also specified.

Compound	NOI	ΔH_d	E_g	ν_{JT}	P	θ_m	Type
$InWV_3S_8$	4	0.09	0.58	5.83	0.56	65.0	II
$AlCrV_3Se_8$	8	3.17	0.19	3.77	1.87	56.4	II
$InMo_4Se_8$	14	-0.69	0.62	4.55	1.08	63.4	I

$InWMo_3Se_8$	19	-0.99	0.63	4.43	0.24	63.8	I
$InCrV_3S_8$	20	2.59	0.40	4.75	0.28	56.6	II
$AlCrV_3S_8$	21	2.63	0.39	5.81	1.02	57.0	II
$InCrV_3Se_8$	25	3.10	0.22	3.45	0.58	56.0	II
$InTaMo_3Se_8$	28	-0.88	0.62	4.25	1.38	54.8	II
$AlTaV_3Se_8$	38	0.56	0.56	3.90	0.15	57.3	I
AlV_4Se_8	47	1.06	0.46	4.08	2.80	54.9	I
$InNbMo_3Se_8$	49	-0.66	0.59	4.44	0.75	55.2	II
GaV_4Se_8	53	1.18	0.44	4.09	2.37	55.0	I

5.4 Summary

Although we have seen an increasing emphasis on using Bayesian optimization for materials design, previous work relied heavily upon handcrafted features, which is a challenging task, or single objective optimization. The former usually requires either knowledge of influential features based on theory and literature or large datasets to perform sensitivity analysis and correlation analysis to identify features that influence properties of interest. In the lacunar spinel MIT materials design, the scientific community is limited by chemical intuition as well as large datasets to identify appropriate features. This hinders the application of traditional BO implementations for MIT design. The propensity to use features arises mainly due to a lack of accurate and efficient machine learning methods to model categorical inputs. Here we showed LVGP can circumvent feature identification by directly modelling elements as categorical variables. The mapping of the categorical variables into low-dimensional quantitative latent variables provides an inherent ordering for the categories and physics-based dimensionality reduction. Like conventional Gaussian process models, the LVGP model provides uncertainty quantification, which is crucial

for employing the BO strategy for material composition optimization. LVGP enables featureless learning and subsequently featureless BO, making it a generic step forward in machine learning and materials design.

The AOE we demonstrated is theoretically more efficient than evolutionary algorithms for identifying the Pareto frontier in a complex, combinational design space. Although designing materials under a single criterion is more efficient, such efforts may not meet the requirements of deployment. For lacunar spinels investigated here, maximizing E_g exclusively leads to an unstable composition while maximizing ΔH_d exclusively leads to a composition with a small bandgap. In contrast, MOBO identifies the Pareto front to delineate the trade-off between materials properties and allows the designer to choose compositions for detailed study. In this context, the need to perform more iterations of MOBO within the AOE is justified. Indeed, it is typically not the sole goal to find all Pareto front designs, but rather to identify the best candidates within a limited research budget. The AOE clearly provides an efficient way to minimize the effort towards a better design by suggesting the next experimental design.

Similar to forward materials design demonstrated here, inverse materials design [148] can be cast as an optimization problem and tackled via the AOE framework. Although forward design is achieved with the objective of maximizing the desired properties, inverse design can be accomplished by redefining the objective as the minimization of the difference between the predicted and target properties. The design space, i.e., the choice of admissible elements, must be defined appropriately to ensure the target properties are achieved. To that end, our work advances materials innovation for forward and inverse design of both inorganic (as shown herein) and organic materials, such as identification of new quantum materials, design of protein sequence in

biomaterials, and monomer sequence in polymeric materials. It is particularly useful when data availability and physical understanding of the target materials system is limited at either the atomic or microstructural scale. This methodology could be further extended to mixed-variable optimization problems, e.g., co-design of composition and chemical stoichiometry through doping, which we are now actively developing.

6 Descriptor Aided Bayesian Optimization for Mixed Variable Materials Design

In Chapters 4 and 5, we have shown the effectiveness of Bayesian Optimization for mixed variable material design. Using LVGP to model the design objectives and quantify uncertainty allows us to seamlessly integrate qualitative variables in the design process. As noted previously, the featureless approach to material design enabled by LVGP is an extremely advantageous owing to the general lack of knowledge of all underlying numerical variables – commonly referred as features / descriptors, that distinguish the behavior of levels. However, there are often situations when domain knowledge may help deducing some of the descriptors that influence material property. For example, Balachandran et al. [90] deduced orbital radii of M, A and X-atoms as features for designing the elastic properties of M_2AX compounds. Their decision to use these features were guided by prevalent domain knowledge indicating a fundamental relationship between the electronic charge density and elastic response of materials. Similarly, Herbol et al. [149] utilized relative dielectric constant and density as descriptors for solvent in their effort to identify optimal combinations of perovskite. Although these descriptors only represent a small fraction of all descriptors that influence the behavior of levels, they could be beneficial in BO which relies on a small dataset to estimate the objective behavior over the design space. In this context, it behooves us to examine the benefits of using descriptors when they are available, in conjunction with the LVGP for BO.

In this chapter, we investigate the utility of descriptors under two major themes:

i. For mixed variable optimization problems with low dimensional qualitative variable (~ <15 levels for all qualitative variables), we will investigate whether the inclusion of descriptors can expedite the convergence.

ii. For mixed variable optimization problems with least one high dimensional qualitative variable (at least one qualitative variable >15 levels), we will investigate new methods to leverage the knowledge possessed by descriptors to inform LVGP of the influence of levels which have not been observed. For such problems, initiating BO is computationally expensive since we require the training dataset to contain at least one observation per level per qualitative variable. Consequently, the ability to initiate BO with a small dataset containing a subset of selected levels presents a significant advancement.

6.1 Review of descriptor based Bayesian Optimization

Descriptor based machine learning and material design has been prominent in the age of data-driven material science. The motivation to use descriptors are twofold: (a) descriptors with strong correlation with material properties can lead to machine learning model with high accuracy for small/medium sized datasets and, (b) using descriptors as inputs to machine leaning model can circumvent the design representation issues encountered when dataset contains observations from multiple materials families (for e.g., binary, ternary, quaternary etc.). While the latter is important and justifies further examination, we will limit our discussions to the first motivation in this chapter.

Identifying informative descriptors can be accomplished in two ways: (*i*) domain knowledge provided by experts or collected from literature, (*ii*) employing feature engineering techniques to identify a small set of uncorrelated descriptors using available dataset. Balachandran et al. [90]

used domain knowledge to select orbital radii of M, A and X-atoms as features for designing the elastic properties of M_2AX compounds. Yuan et al. [150] utilized common feature engineering techniques such as Pearson Correlation (to remove correlated descriptors) and Gradient Boosting (for ranking descriptors based on their importance) to identify seven descriptors from a pool of 37 for BO based discovery of piezoelectric. In a similar spirit, Shields et al. [151] used Pearson correlation to prune the set of descriptors for chemical reaction optimization. More examples of feature based BO applied for material design can be found in the review article by Lookman et al. [78].

In some cases, a large number of descriptors are required to fully capture the effects of design variables on the objective. A high-dimensional input subsequently adds to the complexity of modelling the objective with a small dataset. Researchers have devised new methodologies to tackle such scenarios. The COMBO (Common Bayesian Optimization Library) proposed by Ueno et al. [152] suggested using random feature maps to approximate a gaussian process with Bayesian linear model to reduce computational cost. Ling et al. [129] proposed the FUELS (Forrest with Uncertainty Estimates for Learning Sequentially) framework by Random Forrest for tackling high dimensional inputs and presented a method for quantifying uncertainty via jackknife-style variance estimates. They demonstrated the FUELS framework to outperform COMBO for a variety of materials design problems.

Another interesting approach to utilizing descriptors for mixed variable design problems was recently presented by Hase et al. [153] via their GRYFFIN framework. The essence of GRYFFIN is to project each qualitative variable on a simplex and quantify the similarity between levels by measuring the Euclidean distance on the simplex. While the idea of projecting qualitative variable

to a continuous space is analogous to LVGP, GRYFFIN differs due to the fact that it explicitly learns the transformation from descriptor space to the simplex whereas LVGP does the projection implicitly. The learnt projections are then used to predict the optimization objective using a Bayesian Linear Regression model. Three versions of GRYFFIN – Naïve, Static and Dynamic were presented and compared on a variety of benchmark optimization problems. Dynamic GRYFFIN, with its ability to update the projections learnt using Bayesian Neural Network as more data is observed during BO, outperformed other GRYFFIN variants as well as open-source BO packages.

These examples of descriptor based BO for materials design encourage us to examine the benefits of using descriptors in conjunction with LVGP. The dimension of qualitative variable(s), defined here by the number of levels it can assume, will play an important role in determining how the descriptors are used. Thus, we shall study the role of descriptors for low and high-dimensional qualitative variables separately in the following sections.

6.2 Problems with low dimensional qualitative variables

In this section, we examine the benefits of using descriptors to expedite the convergence of BO using LVGP when qualitative variables are low dimensional (~ <15 levels for all qualitative variables). These problems can be solved by current LVGP capabilities, and our goal is to investigate whether inclusion of descriptors is advantageous.

The first question that arises is how to incorporate descriptors for qualitative variables in a LVGP model? As noted in Chapter 2.3, LVGP uses latent variables $\boldsymbol{z}(l_1), \boldsymbol{z}(l_2)..,\boldsymbol{z}(l_m)$ to capture similarity between levels $l_1, l_2, .., l_m$. In theory, the difference between these levels can be completely described by a set of underlying physical descriptors $v_1, v_2 \ldots v_p$. The challenge arises

due to the fact that we seldom know the full set of descriptors. For several problems in material design, we only have knowledge of a handful of important descriptors based on domain knowledge. Thus, for a practical approach, we desire the ability to incorporate the known descriptors as well as to use latent variables to estimate the effects of unknown descriptors on the levels of each qualitative variable. One way to accomplish this task is to include the descriptors as an auxiliary numerical input (along with any quantitative variables) to the LVGP model and constrain the levels to be closer to each other in the latent space. The constraint is applied in the form of a penalty, which is formulated as the sum of all inter-level distance in the latent space. The motivation for this penalty is simple – if the incorporated descriptors explain the similarities between levels, then there are no effects to be captured in the latent space and consequently the inter-level distance must be small. Note that inter-level distances $\left|\left|\boldsymbol{z}^{(i)}(t_i) - \boldsymbol{z}^{(i)}(t_i')\right|\right|_2$ is used in the computation of correlation matrix (Eq. (2-4)). The penalty term is subtracted from the loglikelihood term and minimized using optimization routines during model fitting:

$$\mathcal{L}(\mu, \sigma^2, \boldsymbol{\varphi}, \boldsymbol{Z}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln|\boldsymbol{r}(\boldsymbol{\varphi}, \boldsymbol{Z})| - \frac{1}{2\sigma^2}(\boldsymbol{y} - \mu\boldsymbol{1})^T\boldsymbol{r}^{-1}(\boldsymbol{\varphi}, \boldsymbol{Z})(\boldsymbol{y} - \mu\boldsymbol{1}) \\ - n * \lambda\sum_{j=1}^{q}\left(\sum_{i=1}^{m_j}\sum_{k=i+1}^{m_j}\left|\left|\boldsymbol{z}^i(t_j) - \boldsymbol{z}^k(t_j)\right|\right|_2\right). \tag{6-1}$$

The hyperparameter λ determined the severity of penalty and can have values ranging from 0 (no penalty) to 1 (severe penalty). We recommend using Leave-One-Out-Cross Validation (LOOCV) to identify the optimal value of λ. It is important to include λ=0 in the LOOCV search for optimal λ since there may be situations when the descriptors do not provide meaningful information about

the similarity of levels and thus the latent variables must be used to learn them. In such case, application of a penalty is counterproductive and models with λ=0 must be the most accurate.

6.2.1 Numerical tests

In this section, we shall use a set of optimization test functions to examine the benefits of using descriptors as auxiliary inputs to LVGP for BO. For a preliminary study, we use two 2D functions – Branin Hoo function (Eq.(6-2)) and Goldstein Price function (Eq. (6-7)). In their original form, both functions have two quantitative variables. We modify both functions such that one of the variables can only assume discrete values and thus transformed to a qualitative variable. It can assume four and five levels for Branin Hoo and Goldstein Price function, respectively.

$$f_{BH}(x_1,t) = \left(t - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10 \qquad (6\text{-}2)$$
$$-5 \le x_1 \le 10,\ t \in \{0,5,10,15\}$$

$$f_{GP}(x_1,t) = [1 + (x_1 + t + 1)^2(19 - 14x_1 + 3x_1^2 - 14t + 6x_1t + 3t^2)][10 + (2x_1 - 3t)^2(18 - 32x_1 + 12x_1^2 + 48t - 36x_1t + 27t^2)] \qquad (6\text{-}3)$$
$$-2 \le x_1 \le 2, \mathrm{t} \in \{-2,-1,0,1,2\}$$

As pointed out previously, selecting the appropriate value of penalty parameter λ is not possible since its rarely known how well the descriptors describe differences between levels. Thus, at each BO iteration, we fit models with $\lambda = 0,0.01,0.1,1$ and then select the best model using LOOCV. We use this wide range of values for λ to account for cases when descriptors are not informative ($\lambda = 0$), descriptors are highly informative ($\lambda = 1$) and all intermediate scenarios. In addition to testing the proposed method with LVGP regularization, we consider two additional approaches. First, we consider the case when no descriptors are available for the qualitative variable, and we

use the original LVGP model (no regularization). Second, we consider the situation where a conventional GP model (with quantitative inputs) is used along with the descriptors and the qualitative variable is neglected. This approach assumes that the descriptors accurately explains the differences between levels of qualitative variable and thus, are sufficient to train an accurate model. For the two test functions studied here, we know the true descriptors and thus consider the approach with conventional GP model with descriptors as the benchmark. In Figure 6-1 we plot the history of objective function values observed throughout BO. We notice that all three approaches converge to the optimum solution at approximately the iteration number. A significant observation to be drawn here is that the performance of BO with LVGP model (does not include descriptors) is as good as BO with GP model which we consider as the benchmark. This indicates that there is little room for improvement in BO with the inclusion of descriptor.

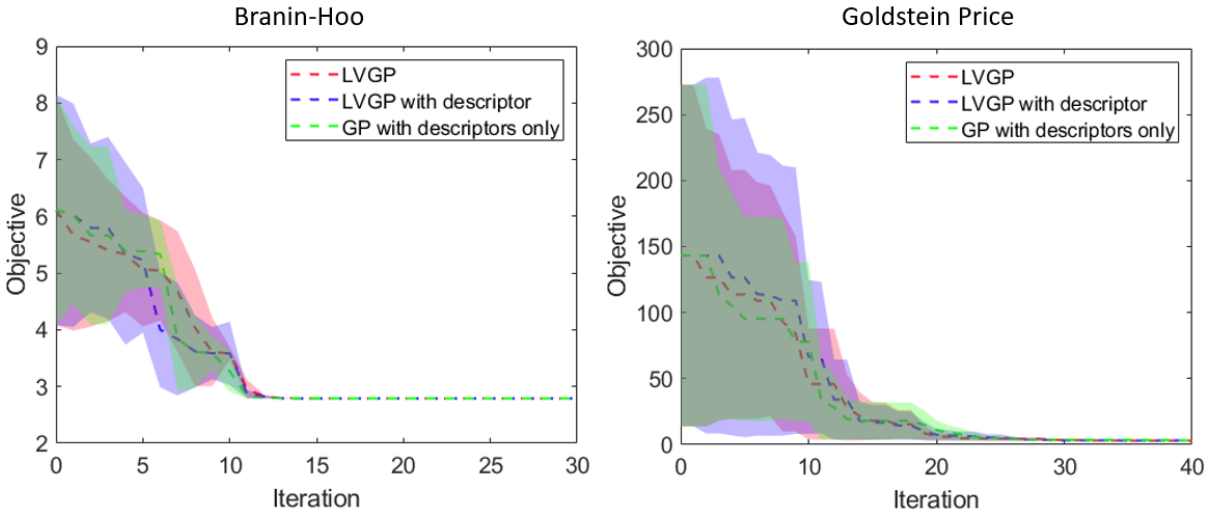


Figure 6-1: Comparing performance of three different approaches in BO for Branin-Hoo and Goldstein Price functions. Red curve represents the original LVGP model (no descriptors included), blue curve represents penalized LVGP model with descriptors and green curve shows

conventional GP model with descriptors only. The dashed lines show the median and the corresponding envelop represents median absolute deviation.

To examine whether the dimensionality of descriptors influences convergence of these three modelling approaches, we use the n-dimensional Levy function:

$$f_{Levy}(\boldsymbol{x}) = \sin^2(\pi w_1) + \sum_{i=1}^{3}(w_i - 1)^2[1 + 10\sin^2(\pi w_i + 1)] + (w_3 - 1)^2[1 + \sin^2(2\pi w_3)] \quad (6\text{-}4)$$

$$w_i = 1 + \frac{x_i - 1}{4}; i = 1,2..n$$

We consider two versions of the Levy function – 7D and 12D. In both versions, the first two dimensions are treated as two independent quantitative variables while the remaining dimensions are assigned to a qualitative variable with five levels. Thus, for the Levy 7D and 12D functions, the qualitative variable has five and ten underlying descriptors, respectively. The descriptors for each level were assigned using a Latin Hypercube sampling. Based on the convergence history shown in Figure 6-2, we notice no significant difference in the convergence history of the three approaches considered here.

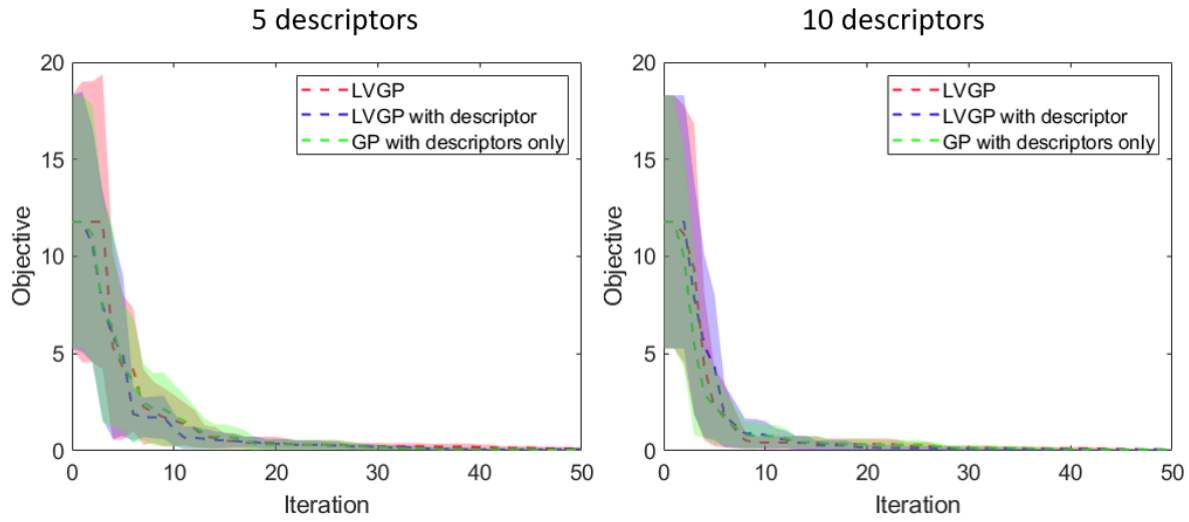


Figure 6-2: Comparing performance of three modelling approaches in BO for Levy function with 5 and 10 descriptors for qualitative variable. Red curve represents the original LVGP model (no descriptors included), blue curve represents penalized LVGP model with descriptors and green curve shows conventional GP model with descriptors only. The dashed lines show the median and the corresponding envelop represents median absolute deviation.

Finally, we compare the three approaches for a materials design application inspired by Balachandran et al. [90]. The objective is to identify M_2AX compounds with optimal mechanical properties. There are three qualitative design variables corresponding to the choice of chemical elements that could occupy the M, A and X sites in the M_2AX crystal. There are 10,12 and two levels for the M, A and X qualitative variables. The authors of the dataset recommend s, p and d-orbital radii of M site and s, p orbital radii of A & X sites as descriptors based on their knowledge of this material system. Out of the 240 possible combinations, only 223 have non-negative mechanical properties are part of the design space. Three mechanical properties considered here are the Bulk, Shear and Young's modulus and BO was performed to maximize each property individually as shown in

Figure 6-3. For the Bulk and Shear Modulus optimization, we notice that the LVGP approach which does not involve use of descriptors performs better than the approaches that use descriptors. The conventional GP model using only the descriptors performs the worst suggesting that the descriptors do not capture all differences between levels. However, for the optimization of Young's modulus, we notice the two approaches using descriptors perform better than the original LVGP model, with the penalized LVGP model with descriptors performing the best. These results indicate that the atomic orbital descriptors do indeed provide valuable information for Young's modulus.

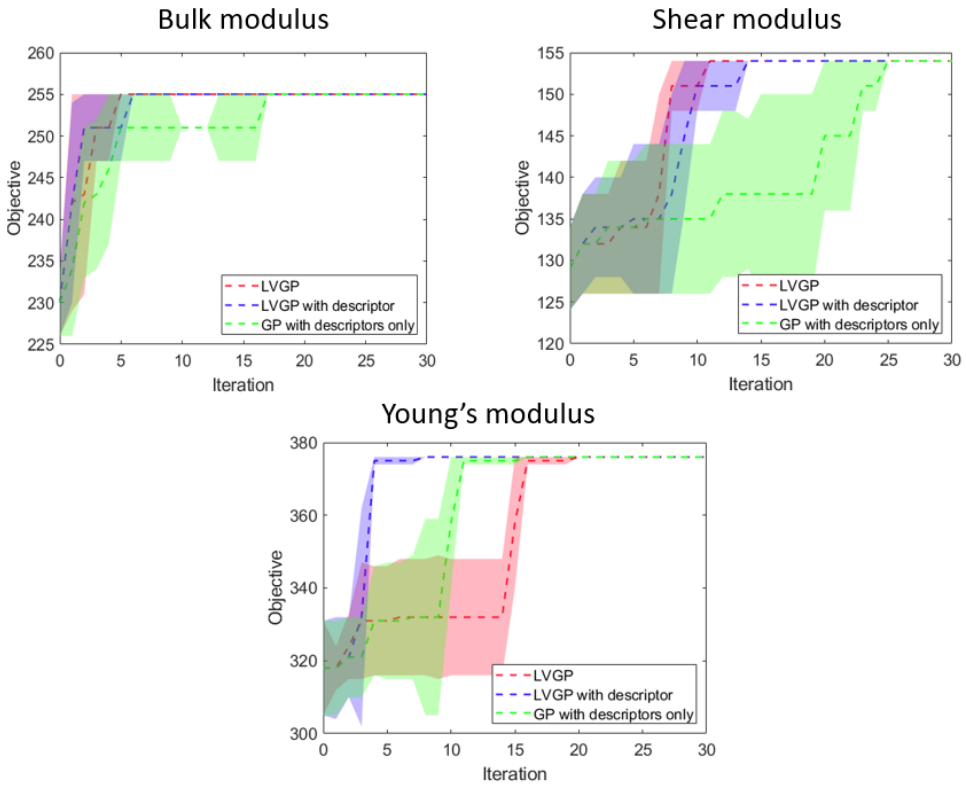


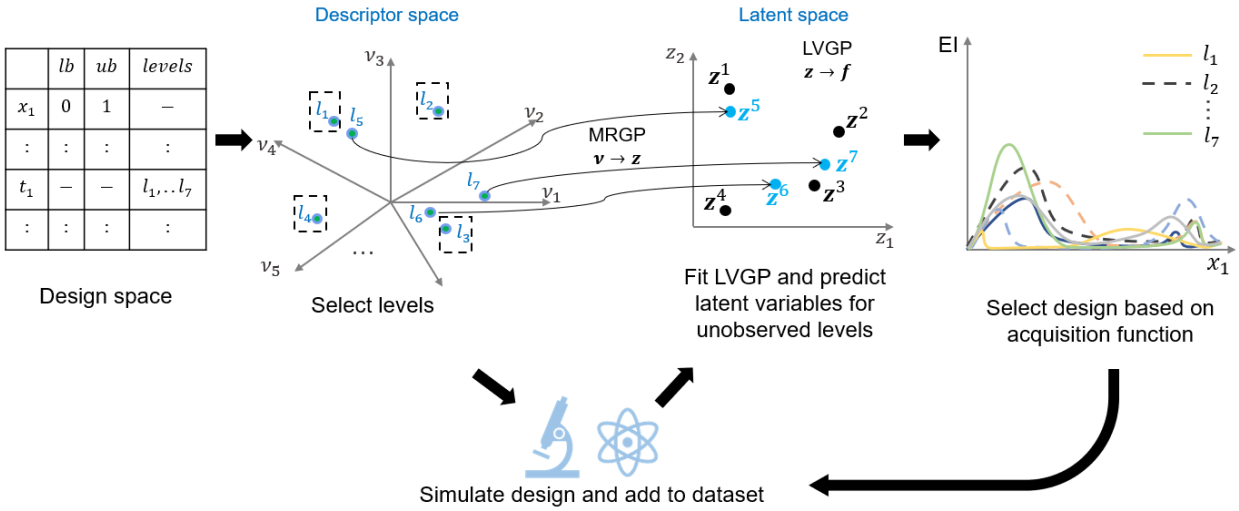
Figure 6-3: Mechanical property optimization history for MAX dataset. The dashed lines show the median and the corresponding envelop represents median absolute deviation.

6.3 Problems with high dimensional qualitative variable

In this section, we examine the use of descriptors for problems where at least one qualitative variable is high dimensional i.e., has a large number of levels. We loosely define large number of levels in this context as more than 20 levels. We encounter unique challenges in performing BO when one or more qualitative variable is high dimensional. A large dataset, containing at least one observation for every level of each qualitative variable, is required to initiate BO. It follows from the fact that LVGP requires at least one observation for every level to estimate its corresponding latent variables. This requirement significantly increases the computational overhead of BO.

Here we propose a new methodology that leverages domain knowledge to tackle high dimensional qualitative variables. Based on the observation that latent space estimated by LVGP model has an inherent structure with levels closer to each other having similar effect on response, we argue that that only a subset of levels are sufficient to initiate BO and that the effect of unobserved levels can be deduced based on their similarity w.r.t the chosen levels. Shown in Figure 6-4, our modified BO approach differs from the conventional approach in two significant ways. First, we use the descriptors to select a diverse subset of levels which will be evaluated to start BO. Second, at every iteration we identify the levels that have not been observed yet and predict their latent variables using a Multi-response Gaussian Process (MRGP) model. In essence, the MRGP model learns the mapping from descriptor space to latent space ($g: v \rightarrow z$) and helps us account for the effects of unobserved levels. Note that latent variables for levels with at least one observation in the dataset will be estimated during LVGP model fitting procedure and will be used to train the MRGP model. Thus, latent variable prediction is only performed for levels which have not been observed yet. This is a critical step in our methodology since it allows us to account for

all feasible levels for a qualitative variable and evaluating the acquisition function for a previously unobserved level.



	lb	ub	levels
x_1	0	1	–
:	:	:	:
t_1	–	–	$l_1,..l_7$
:	:	:	:

Figure 6-4: Bayesian Optimization framework for high dimensional qualitative variables.

6.3.1 Selecting a diverse subset of levels

In this section, we describe the procedure used to select a diverse subset of levels required for initiating BO. In this context, diversity is defined w.r.t the set of descriptors $\boldsymbol{\nu}$ $(\nu_1, \nu_2, \ldots, \nu_p)$ that delineate the differences between influence of levels on the response. Since these descriptors are essential numerical values, the problem of selecting a diverse subset of levels is analogous to the task of design of experiments i.e., selecting a diverse set of designs in a p-dimensional space. Here we use the concept of D-optimality for selection. The procedure starts by calculating a similarity matrix S for a set of levels such that:

$$S_{i,j} = \exp\left(-0.5 * d_{i,j}\right), \tag{6-5}$$

where $d_{i,j}$ is Euclidean distance between levels i and j measured in the descriptor space. The similarity matrix S is positive semi-definite, and its determinant represents the volume spanned by

the set of levels in the descriptor space. Thus, our task of selecting a diverse subset of levels B from the set of all feasible levels A can be accomplished via an iterative process where the level leading to the largest determinant of S is selected. This method of level selection of a qualitative variable is summarized as Algorithm 1 below. When there are multiple qualitative inputs, this method is applied for each qualitative variable independently.

Algorithm 1: Selection levels for a qualitative variable t using descriptors

---

A: set of all levels for qualitative variable t
B: set of selected levels qualitative variable t
User Inputs: Number of levels to select N_{select}, Descriptors $\boldsymbol{\nu} = (\nu_1, \nu_2.., \nu_p)$

Initialize B as empty set : $B = \emptyset$

Add a randomly selected level l from A to B: $B = B \cup \{l\}$

Remove l from A

---

for $i = 1 : N_{select} - 1$

  for $j = 1: size(A)$

    C: B $\cup A_j$

    Calculate interlevel distance $d_{l_m l_n} = \sqrt{\sum_{k=1}^{p} \left(\nu_k^{l_m} - \nu_k^{l_n}\right)^2}$ for all levels in C

    Calculate similarity matrix as $S_{l_m l_n} = exp\ (-0.5 * d_{l_m l_n})$

    $D_j = det(S)$

  end

  Select level l_{select} with largest $\boldsymbol{D}$
  Add selected level to B (B: $B \cup l_{select}$)
  Remove selected level from A ($A: A - l_{select}$)

end

Figure 6-5 illustrates the algorithm for selection of levels based on a two-dimensional descriptor space. The task is to identify a diverse set of 10 levels out of a possible 111 levels. Each level can be described by two descriptors ν_1 and ν_2 which are numerical values spanning [0,1] and used for

evaluating the similarity matrix S in Algorithm 1. Based on the sequence of level selection, it is evident that the algorithm selects levels that are distant for those already selected. While the algorithm is initiated by a randomly selected level which influences subsequent level selection, we note that the levels on the extremities of the descriptor space have a higher probability of being selected.

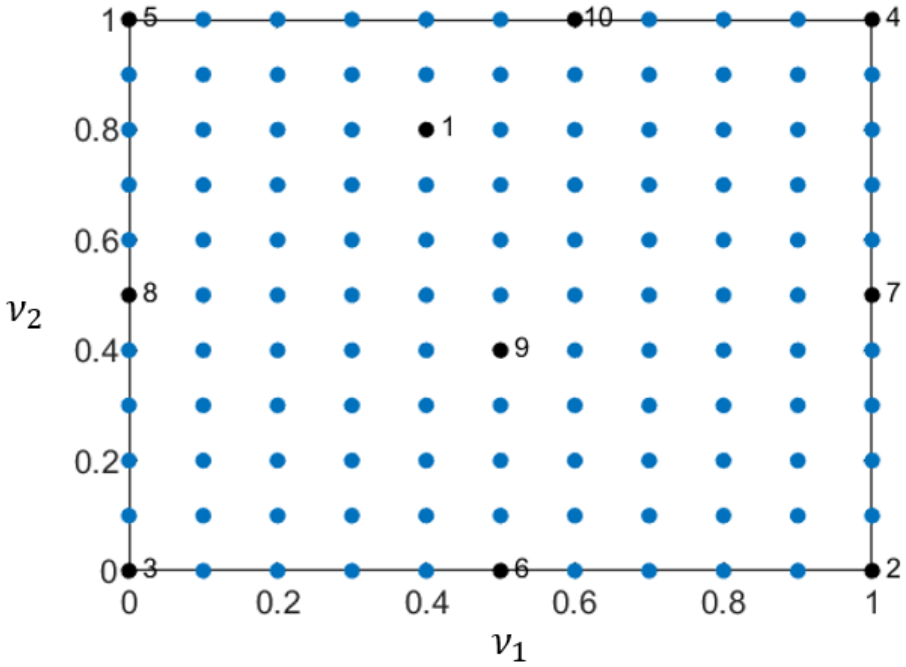


Figure 6-5: An illustration of level selection using descriptors. Out of 111 feasible levels (blue dots) described using two descriptors ν_1 and ν_2, 10 are selected to form a diverse subset (black dots). Each selected level is associated with a number indicating the order of selection.

6.3.2 Descriptor aided latent variable prediction

BO is initiated by collecting observations (experiments / simulations) for the selected levels. These observations are used in LVGP model fitting, and their latent variables are estimated in the process. What latent variable values should we assign to the level that were not selected and hence are unobserved? These levels are still part of design space and must be considered for exploration

in the subsequent BO iterations. A naïve approach would be to assign these levels random latent variable values by drawing them from, for e.g., the uniform distribution U

$$\boldsymbol{z^m} \sim U(-3,3), \tag{6-6}$$

These specific values for bounds of the distribution correspond to the optimization bounds assigned to latent variables during LVGP model fitting. This approach draws upon the ideology that we do not possess any knowledge about the behavior of levels. However, the descriptors used in level selection represents knowledge regarding the behavior of levels i.e., their similarities / dissimilarities. This information can be leveraged to estimate latent variable for levels lacking data for we believe that levels with similar values of descriptors (closer to each other in the descriptor space) will have similar influence on the material property. One approach of utilizing this knowledge is to learn the mapping from the descriptor space to the latent space of LVGP model. Since the latent space is two-dimensional, we seek to use a Multi-response Gaussian Process model to learn this mapping :

$$\boldsymbol{z^m} = G_2\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \otimes r\left(\boldsymbol{\nu}, \boldsymbol{\nu}'\right)\right), \tag{6-7}$$

where G_2 is a two dimensional Gaussian Process with mean $\boldsymbol{\mu} = [\mu_1, \mu_2]$. The covariance function of an MRGP involves the Kronecker product $\otimes$ between the correlation function, here chosen to be gaussian $r\left(\boldsymbol{\nu}, \boldsymbol{\nu}'\right) = \exp\left(-\sum_{i=1}^{p} 10^{\omega_i}(\nu_i - \nu_i')^2\right)$, and a 2x2 symmetric positive definite matrix $\boldsymbol{\Sigma}$. Specifically, the diagonal and off-diagonal entries of $\boldsymbol{\Sigma}$ capture the marginal variances and covariances between outputs respectively. Analogous to the single response GP model, the additional hyperparameters introduced by $\boldsymbol{\Sigma}$ are estimated during model fitting using maximum likelihood estimation. Thus, the role of MRGP is to learn the mapping from descriptor to latent

space using the observed levels and subsequently predict the latent variables $\boldsymbol{z^m}$ for the unobserved levels. These predicted latent variables are used as proxies for unobserved levels to calculate the acquisition function and gauge the benefit of sampling them in future objective evaluations. This method, in essence, uses the knowledge possessed by descriptors to augment the LVGP model by predicting the latent variables for levels lacking data. Consequently, we denote this approach as Descriptor Augment LVGP.

Figure 6-6 compares the random and descriptor augmented LVGP approaches for Branin-Hoo function which has a qualitative variable with four level as defined in Eq. (6-2). In a training dataset of 60 observation (15 per level), all observation involving level 2 was excluded from the dataset used to train a LVGP model. Thus, LVGP identified latent variables for levels 1, 3 and 4 (black dots) while level 2 was assigned latent variables using the latent space interpolation (blue star) and random (red star). We notice that descriptor augmented technique yields a rather accurate representation of the effects of level 2 while random assignment is highly inaccurate. The magnitude of the error bars suggests that uncertainty is low in the regions of latent space closer to existing levels, and it increases as we move away from them. This observation is akin to a conventional GP model where the model uncertainty increases as we move away observed data points.

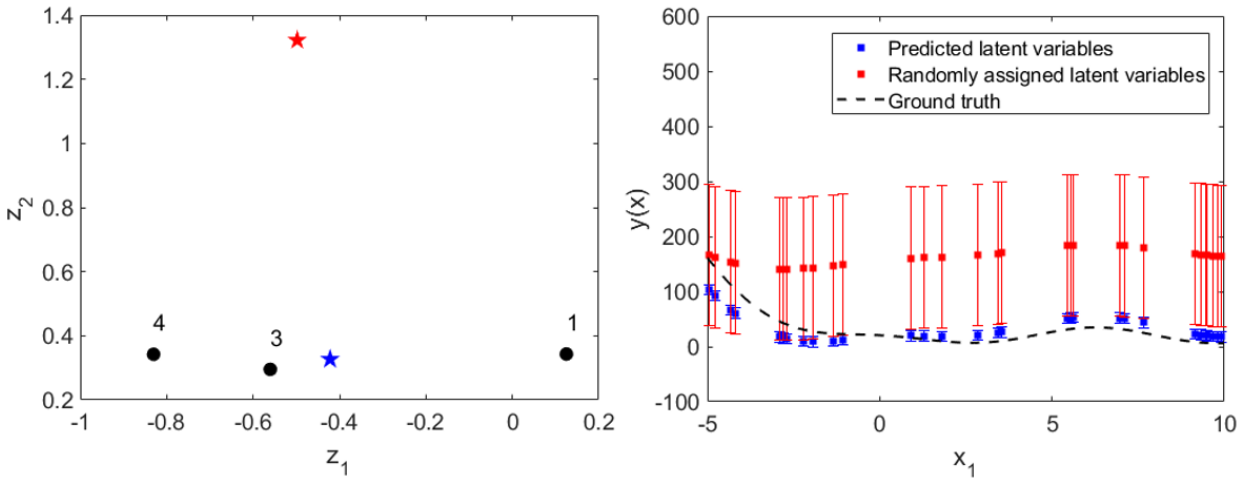


Figure 6-6: Comparing latent variable assignment techniques (left figure) and its impact the predictive performance (right figure) for Branin-Hoo function. The error bars show the predicted standard deviation.

6.3.3 Mathematical Benchmarks

In this section, we test the efficacy of our approach for test functions specifically designed to have a high dimensional qualitative variable. We revisit the Levy function described in Eq. (6-4). Here, we consider 4D Levy function with the third and fourth dimension mapped to a single qualitative variable. Figure 6-7 shows the uniform distribution of levels in these two dimensional descriptor spaces defined by x_3 and x_4 when there are 26 and 56 levels. To perform BO using the conventional LVGP model (Vanilla LVGP in Figure 6-7), we require 26 and 56 observations (i.e., one observation per levels) for these scenarios respectively. The excessive computational cost incurred due to linear scaling of dataset size with the number of levels emphasizes the limitation of LVGP method for BO in presence of high dimensional qualitative variables. On the other hand, modifications to LVGP described in Sec. 6.3.2 enable initiation of BO with a smaller dataset containing seven levels selected via Algorithm 1. The values for quantitative variables x_1 and x_2 are generated using Latin hypercube design. We compare the performance of two proposed

approach – LVGP-Descriptor Augmented and LVGP-random against the conventional GP model that uses the numerical descriptors as input. When the true underlying descriptors for the qualitative variables are known, as in this example, we expect the GP model using descriptors to have superior performance and serve as a benchmark. The convergence history shown in Figure 6-7 corroborates this belief, as the GP approach has better convergence characteristics followed by Descriptor Augmented LVGP approach. In fact, for the scenario with 26 levels, Descriptor Augmented LVGP and GP approaches have comparable performance beyond 20 iterations. The LVGP-Random approach, where latent variables are sampled from a uniform distribution without any consideration for descriptor-based similarity, underperform as compared to the other two approaches initially.

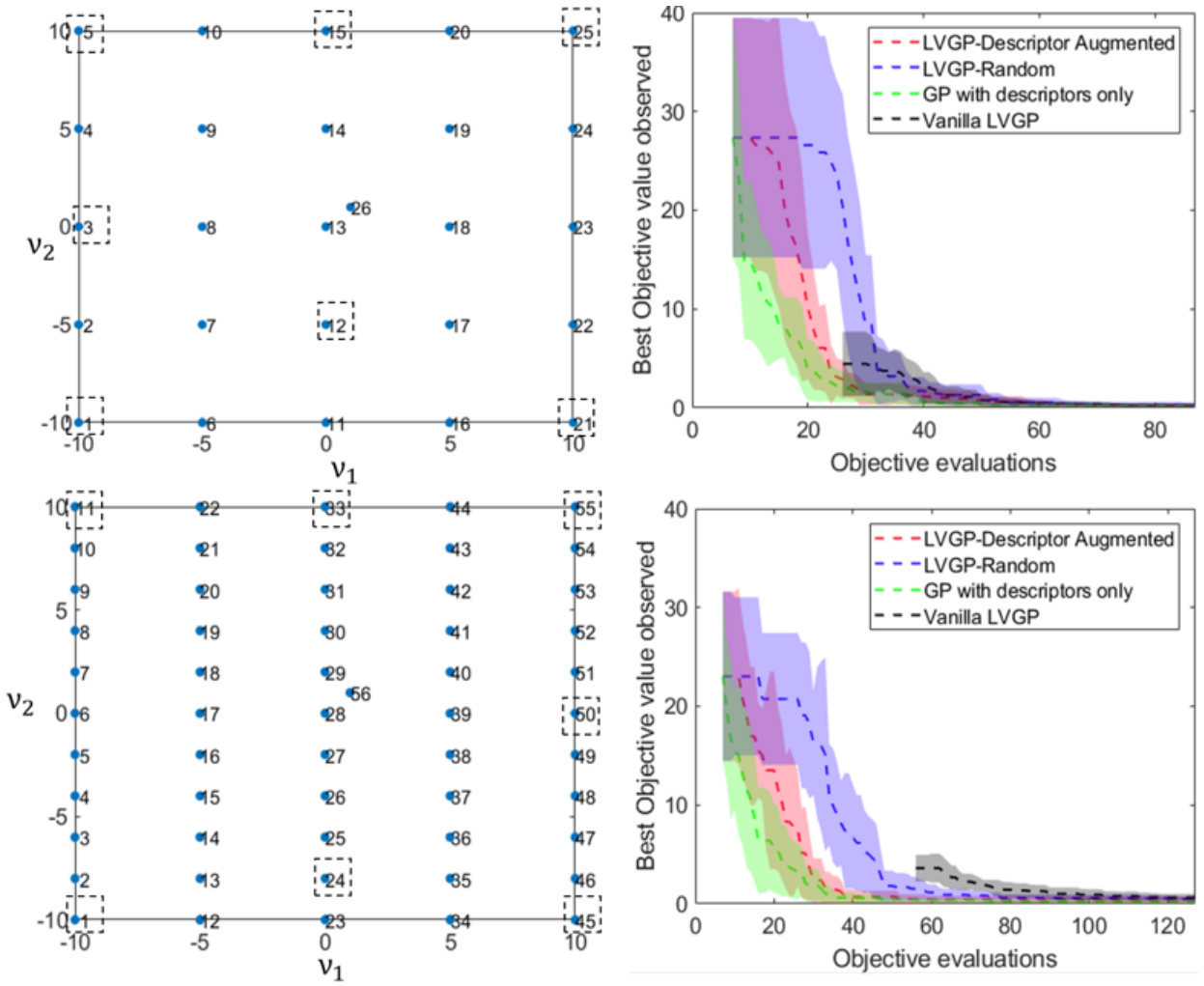


Figure 6-7: Comparing performance of different approaches for Bayesian Optimization of 4D Levy function when qualitative variable has 26 (first row) and 56 levels (second row). The scatter plots show the distribution of levels and the black boxes with dashed lines indicate levels selected in one replicate. The optimization history is shown using median and median absolute deviation computed over 30 replicates.

To better understand the differences between these three competing modelling approaches, we now examine the designs sampled in the first 20 iterations of BO for each model. We consider this initial BO phase for examination as it will be affected the most by the assumptions made by

models. Figure 6-8 shows the sampling history along with contours of Levy 4D function. We observe that levels in the vicinity of origin typically have lower objective values as compared to the to the other levels and are more likely to be sampled during BO. This is corroborated by the sampling histories for GP and LVGP approaches leading to the significant improvements in objective values observed in BO. The GP model using the true underlying descriptors strongly favors level 26 which contains the optimum. On the other hand, the LVGP-Random approach samples all levels uniformly. This could be due to the high uncertainty associated with randomly assigned levels especially when they are located far away for the observed levels (depicted in Figure 6-6).

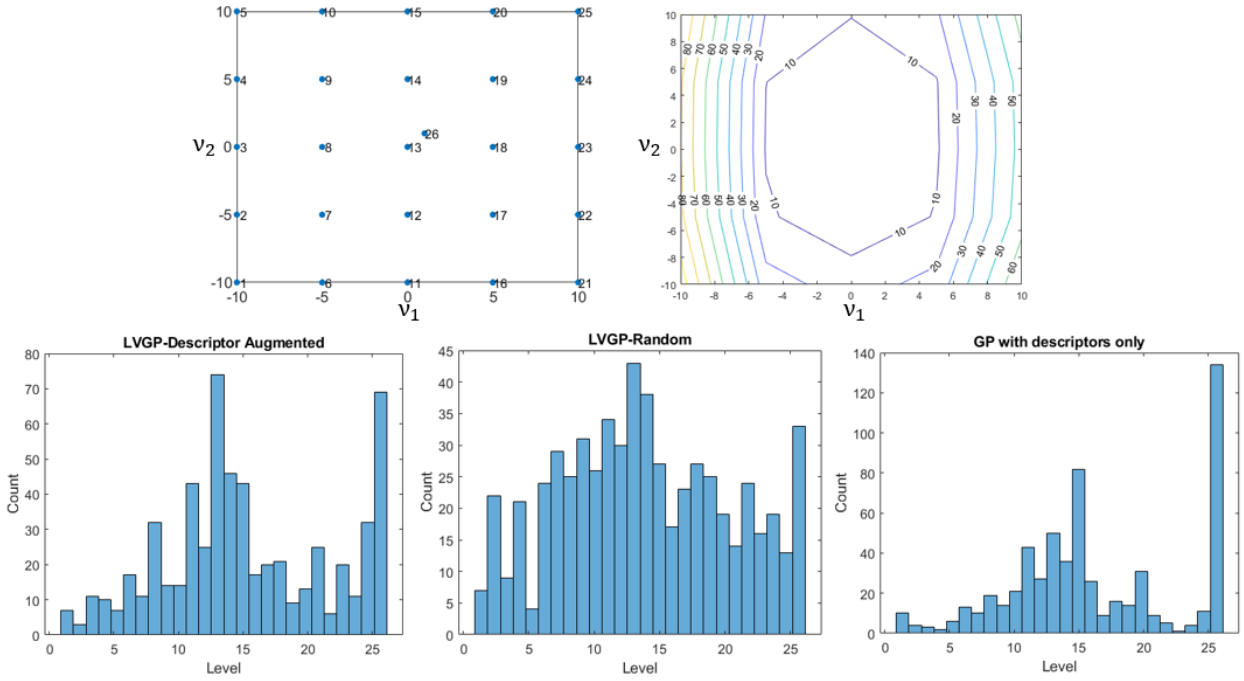


Figure 6-8: Upper row: distribution of levels in the two dimensional descriptor space (left) and contour of objective function in this space (right). For contour plot, quantitative variables were set to their mean values. Lower row: Sampling history for qualitative variable in Levy 4D function

during the first 20 iterations of BO. Histograms show the consolidated history over 30 replicates for each model.

Effect of incorrect descriptor knowledge

In the above comparisons for Levy function, we assumed to have perfect knowledge regarding the two underlying descriptors (x_3 and x_4) for qualitative variables. However, this is far from the scenario encountered in practice where we only have knowledge of some descriptors, or the descriptors are incorrect and hence not informative at all. To simulate the latter scenario, we reconsider the Levy function with 26 levels with the descriptors assigned randomly for each level as shown in the lower scatter plot in Figure 6-9. For example, while the true descriptors for level 13 are {0,0}, it is assigned {-10,-5}. This discrepancy between the true and assigned descriptors has a significant impact on the GP model which is dependent on the descriptors for modelling the objective function. On the other hand, Descriptor augmented LVGP utilizes descriptors only for unobserved levels. Once a level is observed, it becomes part of the LVGP model fitting and does not depend on descriptors. Consequently, we observe that its performance is not affected significantly, and it converges to a better design as compared to GP based model.

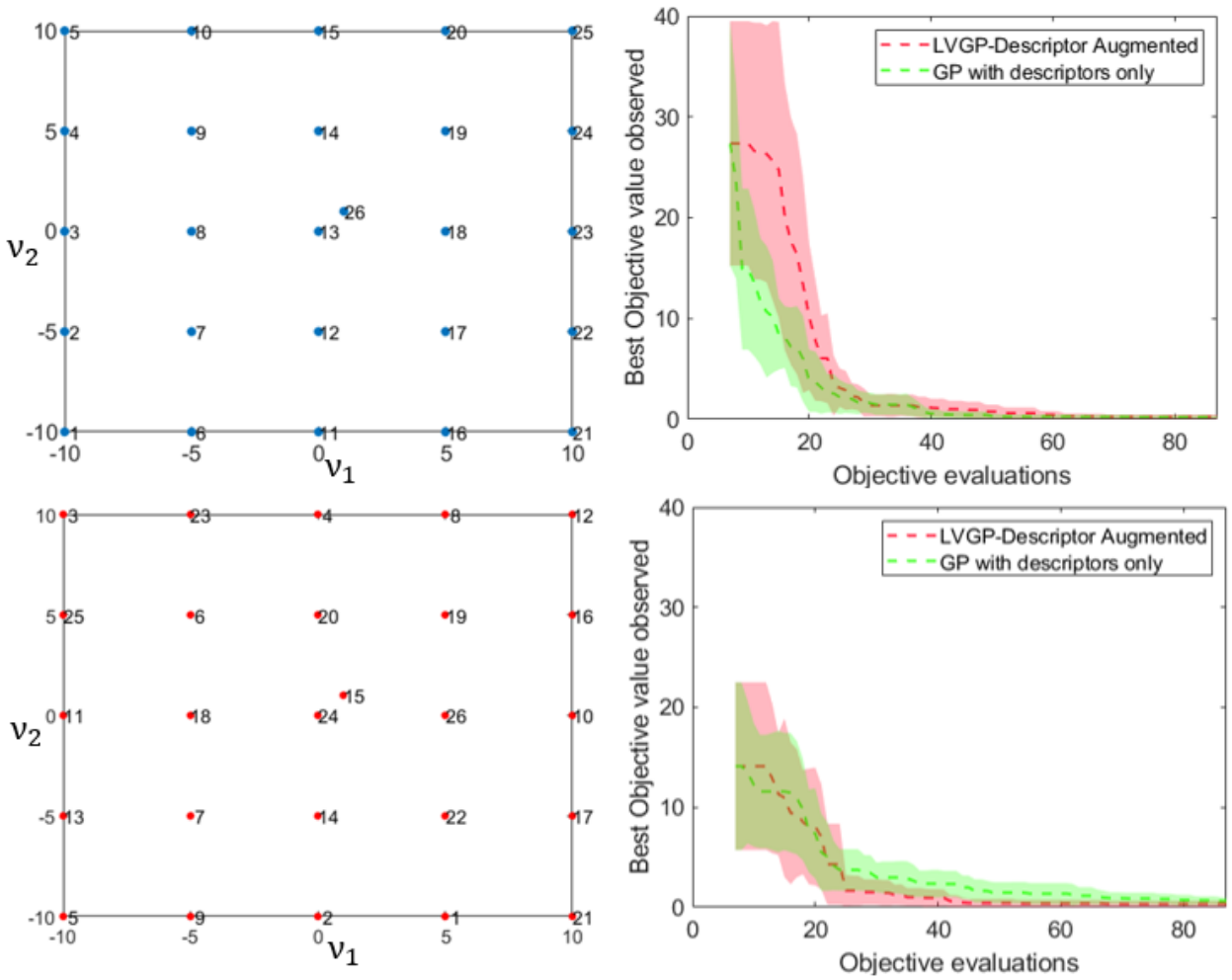


Figure 6-9: Comparing convergence history for BO when the knowledge of descriptors is perfect (first row) and imperfect (second row) across 30 BO replicates. The scatter plots on the left show the distribution of levels in descriptor space. Descriptors were assigned to levels randomly in each replicate of BO to simulate the imperfect knowledge scenario.

Effect of partial descriptor knowledge

Another common situation encountered in practice is that the designer has partial knowledge of descriptors i.e., not all descriptors are known. To study this scenario and the performance of different modelling approaches in BO, we consider the six-dimensional Levy function. The first two dimensions are treated as independent quantitative variables while the other four dimensions

are combined into a single qualitative variable with 26 levels. Hence, the effect of each level can be deduced from its four descriptors $\{\nu_1, \nu_2, \nu_3, \nu_4\}$. The levels were constructed using Latin hypercube design so that they located in different regions of the four dimensional spaces. Based on the Sobol total sensitivity indices listed in Table 6-1, it is evident that the objective function is most and least sensitive to ν_3 and ν_1 respectively. BO was performed by systematically pruning the set of descriptors and its convergence history plotted in Figure 6-10. For the GP model relying entirely on the knowledge of descriptors, we notice the performance deteriorate progressively as more descriptors are dropped with a drastic deterioration when only one descriptor is used. In contrast, the descriptor augmented LVGP approach is only affected significantly when one descriptor is used and shows no change in other scenario. This can be attributed to the fact that LVGP relies on the descriptors only for levels unobserved. Once the level is observed, it's latent variables are estimated during model fitting and is unaffected by descriptors.

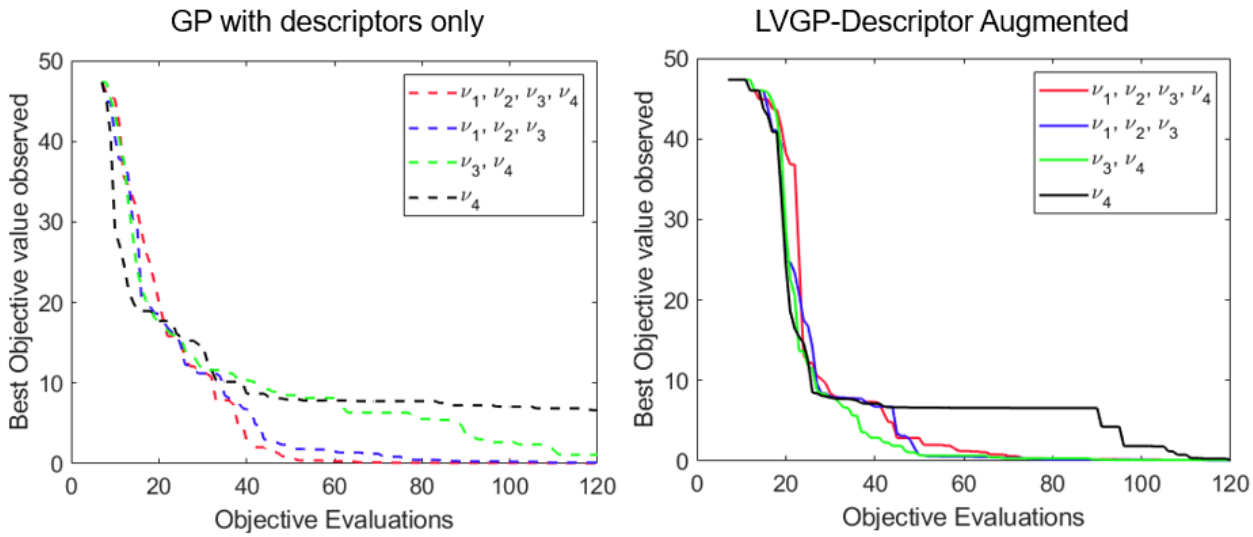


Figure 6-10: Effect of sequential descriptor pruning on BO for Levy 6D function. Lines show median objective values observed across 30 replicates.

Table 6-1: Total sobol sensitivity indices for Levy 6D function

x_1	x_2	v_1	v_2	v_3	v_4
0.198	0.199	0.203	0.179	0.2174	0.007

Varied effect of descriptors on objective

Since the amount of influence asserted by each descriptor on the objective function may vary, the impact of leaving out a descriptor from the pool of all descriptors can be different. To demonstrate this assertion, we again consider the Levy 6D function with one qualitative variable whose effect is determined by four underlying descriptors. We simulate two BO scenarios by only considering descriptors 1 & 2 and 3&4. Figure 6-11 shows the comparative performance of Descriptor augmented LVGP and GP models. We notice that the GP model which relies on descriptors to model the objective, has a significant deterioration in performance in comparison to the descriptor augmented LVGP model. Specifically, we notice that GP model with descriptors 1&2 is impacted more significantly indicating that either descriptor 3 or 4 (or both) have a greater influence on the objective function. To further delineate the sensitivity of these descriptors, we examine the values of roughness (scale) parameters and Sobol sensitivity indices for each descriptor as shown in Table 6-2. Rather than examining the effect of descriptors globally, we specifically examine the effect of descriptors around the global optimum defined by $x_1, x_2 \in [0,2]$. The roughness parameters are derived from the correlation function of a GP trained on 781 observations, 30 observations for each of the 26 levels plus the optimum design. Larger value of roughness parameter for a descriptor indicates effect on objective. Among the four descriptors, the third descriptor has highest

roughness parameter and sensitivity index, indicating its strong influence on the objective in the vicinity of global optimum.

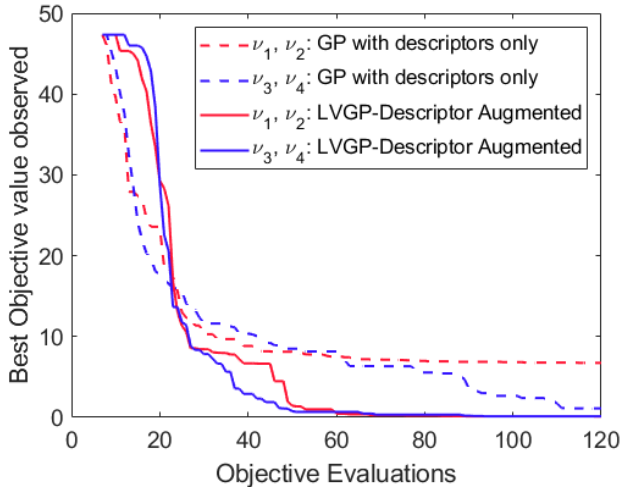


Figure 6-11: Comparison of modelling approaches using a subset of descriptors for Levy 6D function. Lines show median objective values observed across 30 replicates.

Table 6-2: Roughness parameters and Sobol sensitivity index for four descriptors governing effect of qualitative variable in Levy 6D function.

	ν_1	ν_2	ν_3	ν_4
GP Roughness parameters	-0.458	-3.261	0.534	-11.164
Total Sensitivity Index	0.335	0.297	0.360	0.012

6.3.4 Design of ABO_3 Perovskite

We shall next consider a materials design example involving high dimensional qualitative variables. The task is to identify ABO_3 perovskite which are candidate materials for next generation thermochemical water splitting applications and were explored in detail by Emery et al. [154]. In the ABO_3 crystal structure, O represent Oxygen atoms while the A and B crystal sites are occupied by anion and cation respectively. There are 73 elements from the periodic table that

can possibly occupy these sites. Our task is to find the optimal combination of A and B site elements to maximize Stability and minimize Formation energy. The material properties are evaluated using Density Function Theory calculations which constitutes a predominant portion of the computational cost of design process. Thus, it is highly desirable to optimize properties with the least number of simulations.

To initiate BO using a subset of levels, we identify some descriptors for the A and B crystals sites. Since each level represents a chemical element in periodic table, we refer to a pool of commonly used chemical descriptors used to characterize chemical elements. The cheminformatics Matminer was used to extract this feature pool which included a set of eight descriptors – Atomic Number, Atomic Weight, Covalent Radius, Melting Temperature, Mendeleev Number, Electronegativity, Row Number and Number of Valence electrons. Strongly correlated descriptors can lead to poor machine learning models since they obscure the effect of each descriptor. To circumvent this issue, descriptor pairs with strong correlations ($|\rho_{Pearson}| > 0.8$) were identified and selectively pruned. For e.g., Atomic Number and Row Number were removed for the descriptor pool due to their strong correlations with Atomic Weight. Similarly, Mendeleev Number was removed due to its strong correlation with Covalent Radius and Electronegativity. Thus, the final set of descriptors used for the A and B crystals site are – Atomic Weight, Electronegativity, Melting Temperature, Covalent Radius and Number of valence electron. Figure 6-12 shows the optimization history using the three modelling approaches, applied for optimizing Stability and Formation Energy individually. Table 6-3 lists the number of objective evaluations required for the median to match the global optimum for the three models. BO with each model was initiated using nine compounds selected via the procedure outlined in Section 6.3.1. In both cases, model utilizing descriptors

identify the optimum faster than LVGP-Random, indicating the benefits of domain knowledge inclusion in BO. Descriptor Augmented LVGP outperforms GP model with descriptors for Formation energy optimization, in contrast to Stability optimization where the latter performs better. The faster convergence of the two descriptor based models in Stability optimization suggests that the descriptors are informative and aid in capturing the salient effect of each crystal site on Stability.

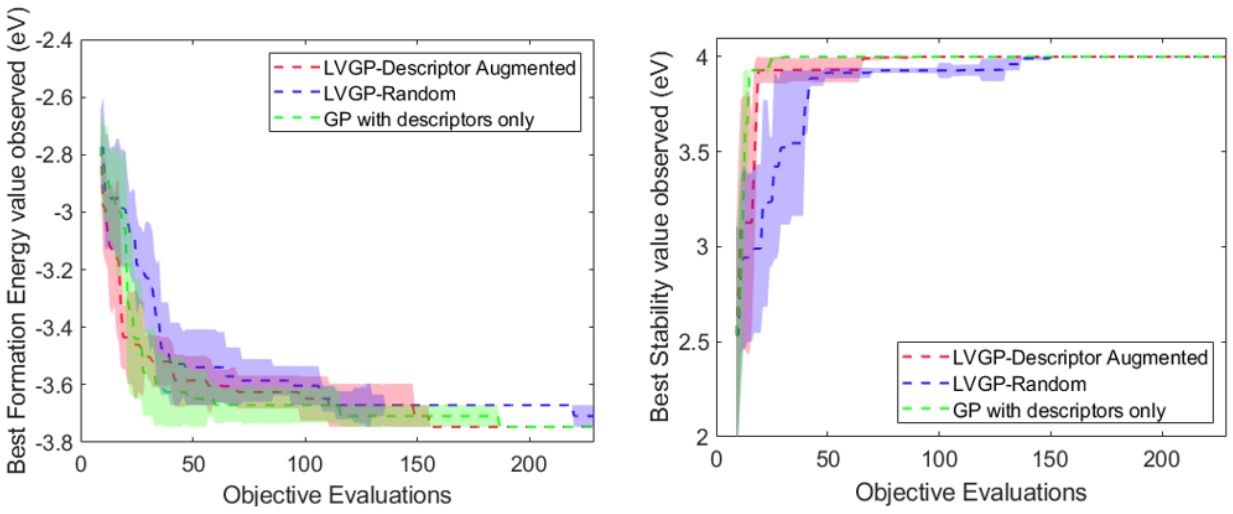


Figure 6-12: BO history for Formation Energy minimization (left) and Stability maximization (right) for ABO_3 perovskites. The dashed lines and envelope represent median and median absolute deviation calculated over 13 replicates.

Table 6-3: Number of objective evaluations required for median objective value to match global optimum

	LVGP – Descriptors Augmented	LVGP – Random	GP with descriptors only
Formation Energy	156	-	187
Stability	31	151	90

6.4 Summary

In contrast to Chapters 4 and 5 which highlighted the benefits of featureless machine learning enabled by LVGP, this chapter was dedicated to study the benefits of incorporating domain knowledge in the form of numerical descriptors. Problems were classified based on the dimensionality of qualitative variables since they necessitate different approaches for fusion of domain knowledge.

For low dimensional qualitative variables, we proposed using descriptors as auxiliary inputs to the LVGP model and addition of a penalty term. The motivation for penalty term was to regularize the latent space when descriptors partially or completely explain the effect of levels on the response. The penalty weight was determined using leave-one-out-cross-validation. Through a combination of numerical test functions and curated material datasets, we observed that our proposed approach did not improve BO convergence to any noticeable extent. In fact, the featureless LVGP approach is as good as, if not better, than the benchmark GP model using the descriptors as inputs. Thus, we recommend utilizing the featureless LVGP modelling approach for best BO performance when the qualitative variables are low dimensional.

For high dimensional qualitative variables with many levels, we realized the large dataset size required to initiate BO using LVGP imposed a severe computational cost. To this end, we proposed descriptor augmented LVGP for BO that starts with a carefully selected subset of levels and leverages descriptors to predict the effect of unobserved levels. Specifically, a Multiresponse GP (MRGP) model was used to learn the mapping between descriptors and latent space to predict the effect of unobserved levels. The decoupled nature of MRGP and LVGP model ensures that effect of observed levels is learnt from the data and not reliant on descriptors. This ability is critical in

many practical scenarios where the designer may have partial or incorrect knowledge of descriptors as demonstrated through a set of test cases. While discussion in this chapter have been limited to single criteria optimization, descriptor augmented LVGP can be extended for multicriteria optimization in the future.

7 Conclusions and Future Work

This chapter summarizes the contributions of this dissertations and describe opportunities for future work.

7.1 Contributions

By addressing design challenges in a diverse set of materials systems, the overarching contribution of this dissertation is to exhibit the ubiquity of data centric design framework presented in Chapter 1. We have developed (*i*) a computational structure – property model to estimate OPVC performance (*ii*) methodology to characterize and reconstruct anisotropic microstructures, (*iii*) data centric polymer nanocomposite design framework by assimilating experimental and simulated data, (*iv*) adaptive optimization engine for composition design of MITs and, (*v*) webtools to provide rapid access to computational MCR tools via NanoMine.

The prominent research contributions made in each chapter are summarized as follows:

a) The salient contributions of Chapter 3 are twofold. First, we developed a novel design evaluation methodology to evaluate the performance of Organic Photovoltaics from its active layer microstructure. This was subsequently used to implement SDF based microstructure design framework to identify optimal, isotopic active layer. Second, we go beyond the traditional isotropic microstructure design formulations to address the design representation limitations plaguing the characterizing and designing anisotropic microstructures in a computationally efficient manner. We developed a novel approach for fast microstructure reconstruction and quantify anisotropy to facilitate design of

anisotropic microstructure. Through a case study, we further demonstrated that anisotropic active layer morphology outperforms its isotropic counterpart.

b) In Chapter 4, we presented a data-centric mixed-variable Bayesian Optimization framework for concurrent design of composition and microstructure, which is inherently a mixed variable optimization problem. The framework integrated modules for (*i*) low dimensional microstructure design representation, (*ii*) calibration and training of design evaluation models using experimental dataset and, (*iii*) BO in tandem with LVGP for mixed variable design synthesis. The efficacy of this framework is exemplified by a polymer nanocomposites case study. Initiated by a nanocomposite database, the 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. Since functional materials must often meet multiple performance criteria, we extended LVGP based BO to multicriteria optimization using the expected maximin improvement acquisition function.

c) In Chapter 5, we developed an adaptive optimization engine which learns directly from elemental compositions to accelerate the co-design of functional electronic materials. Since all design variables involved in this problem are qualitative with varying number of levels, we developed a method to perform design of experiments by discretizing Latin Hypercube design. Our method is particularly helpful in the data and knowledge of materials descriptors is limited. This featureless learning method is readily

applicable to other complex materials design problems. For instance, the same workflow can be utilized for inverse design, where the objective is to minimize the difference to the target value.

d) In Chapter 6, we investigated the design synthesis challenges arising in mixed variable optimization with high dimensional qualitative variable(s). Realizing that the effect of qualitative variables arising in materials design can be described by some underlying chemical descriptors, we proposed a new descriptor augmented LVGP modelling approach for BO. Descriptors are utilized to select a diverse subset of levels to initiate BO as well as predict the latent variables for unobserved level. The decoupled nature of GP models used in the approach enables superior performance in BO even when the knowledge of descriptors is incomplete or imperfect.

e) To provide access to some popular MCR techniques, Appendix 9.1 discusses the current MCR capabilities in NanoMine. A total of eight webtools were created that allow users to binarize images and subsequently use the spatial correlation functions, SDF and physical descriptor based MCR methods. All webtools incorporate some core user-friendly features such as email alerts, multiple image input formats.

7.2 Future Work

The are several open-ended questions and opportunities identified during the formulation this dissertation. Some of them are described below:

a) The design strategy for anisotropic microstructures proposed in Chapter 3 can be extended to design any material systems where anisotropy could potentially enhance performance. While perfect anisotropy was optimal for IPCE maximization, the proposed methodology

is also applicable for multiobjective design formulation where objective behavior could be conflicting w.r.t anisotropy, such as design for high charge conversion and mechanical stiffness. In such cases, the designer will have to select a design from the Pareto front based on his willingness to compromise. Challenges in fabrication of strongly anisotropic morphologies with desired domain size and distribution represent opportunities for future work. In this regard, recent work by Jiali et al. [155] shows that controlled solvent vapor treatment on a highly oriented polyethylene substrate improves crystallization ability and induces strong anisotropy in P3HT films. An alternative approach to induce anisotropy is through application of electric field during annealing, as demonstrated by Dulal et al. [156]. A holistic design approach requires PSP relationships for OPVCs, which is part of our ongoing initiative. Due to expensive, time-consuming nature of P3HT:PCBM film preparation and imaging, one could employ Coarse Grained Molecular Dynamics (CGMD) simulations to mimic film preparation and study effects of key processing parameters such as P3HT chain length, weight ratio of constituents [157], thermal annealing [158] and polydispersity [159]. The stochastic nature of CGMD simulations calls for methods that account of noise and its dependence on simulation inputs, as shown recently by Beek et al. [160, 161]. In future, SDF based microstructure analysis of equilibrated CGMD structures can be used to reveal the SDF pattern and a suitable parameterization. Identifying links between process parameters and SDF parameters will create the process-structure relationship, which can be integrated with structure-performance model used in Chapter 3.3, to complete the PSP chain for OPVC design.

b) Characterization and Reconstruction of multiphase materials remains challenging since one must identify features specific to each phase as well as correlation among different phases. Metal alloys [162] used in a variety of applications and extensively studied recently for additive manufacturing processes [163, 164] are a prominent example of multiphase microstructure systems. Transfer learning approaches that use pretrained convolutional neural networks have been successful in reconstruction from a target [34, 35], but these methods fall short for microstructure design which requires the ability to reconstruct unseen microstructures (no target specified). When a large dataset is available, deep learning techniques such as Generative Adversarial Networks [165] are a viable candidate. Recent works have used this approach for steel microstructure reconstruction [40, 41]. However, these methods do not provide meaningful characterization of microstructures and are not applicable for small datasets.

c) Existing approaches for polymer nanocomposite design has focused on systems with spherical nanoparticles. However, there is a need to investigate the benefits of elongated (i.e., high aspect ratio) nanoparticles which may induce local variation in properties by virtue of their shape. Recent investigations [166, 167] have shown that fillers with high aspect ratio are able to meet targeted electrical property requirements at a lower filler volume fraction than that of spherical particle filled composites. Additionally, development of 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.

d) DFT simulations are one of the most pervasive computational tools to predict structure, energetics, and study atomic scale interactions in material science. Selection of an exchange-correlation functional is one of the critical decisions that DFT practitioners have to make due to its significant impact on the accuracy of calculations. A naïve yet common practice within this field involved trying different exchange-correlations and simulation settings to assess their validity and quantify uncertainty in predicted properties. Given the computationally inefficiency of this approach, Bayesian Error Estimation Functional (BEEF) approach [168, 169] for uncertainty quantification has been developed and widely adopted. BEEF approach leverage ideas from Bayesian statistics to estimate a distribution for the parameters used in exchange-correlations and thus, deduce the distribution of model's predictions. Given the complexity of simulating lacunar spinels (Chapter 5) and the subsequent use of predicted properties in design, it will be beneficial to quantify the uncertainty associated with the DFT simulations and integrate this knowledge in BO.

e) Descriptor assisted BO described in Chapter 6 primarily utilized numerical descriptors (scalar values) to distinguish the effect levels on response. However, descriptors could be formulated in other forms. First, descriptors can be ascribed to the overall design rather than individual qualitative variables. For example, the differences between Metal – Organic – Frameworks compounds is often characterized by geometrical descriptors such as Gravimetric Surface Area, Volumetric Surface Area etc. The methods for level selection and descriptor augmentation must be modified to tackle such problems. Second, data obtained from low fidelity simulations could also contain valuable information regarding the effect of levels. While converting this data into physically interpretable numerical

descriptors as described in Chapter 6 may be challenging, a multi-fidelity BO scheme to extract knowledge from this data could be beneficial.

f) Incorporation of new webtools in NanoMine by introducing standardized data curation workflows, data visualization capability, sophisticated interphase calibration and, FEA tools is necessary to drive the widespread adoption of data centric design methodology in the nanocomposite community. Recently, the Graphical User Interface (GUI) for MCR tools has been updated to allow users to provide additional information regarding images they upload such as scale, resolution. Some additional features that are highly desirable are:

- It will be beneficial to create workflows to process a collection of images (aka batch processing) when each image has a different scale.
- Currently, the MCR webtools act as standalone tools for expedited image processing. However, tapping into NanoMine's ontology enabled knowledge graph requires a framework for storing the information generated by creating protocols for (a) seeking user consent for providing open access to their datasets (images, property measurements, processing conditions) (b) creating a web infrastructure that inserts/deletes/edits the knowledge graph as needed (c) creation of an agent that will periodically process all images uploaded to NanoMine without explicit instructions.
- Similar to other Materials / Cheminformatics packages, it is beneficial to provide an Application Programming Interface (API) in addition to GUI to promote acceptance of NanoMine in the data science community.

This thesis identifies data-centric design methodology as a framework to elucidate the interconnections between the major themes in design engineering and materials science. Discussions on a variety of material systems presented here reinforce the generality of this framework and also shows the unique, system specific challenges encountered in its implementation. We believe the data-centric materials design framework will be the backbone of data driven exploration and development of advancement material systems to meet the challenges facing us in the 21st century.

8 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, 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.

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

[5] A. Jain *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *Apl Materials,* vol. 1, no. 1, p. 011002, 2013.

[6] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)," *Jom,* vol. 65, no. 11, pp. 1501-1509, 2013.

[7] S. Curtarolo *et al.*, "AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations," *Computational Materials Science,* vol. 58, pp. 227-235, 2012.

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

[9] M. Baniassadi, H. Garmestani, D. Li, S. Ahzi, M. Khaleel, and X. Sun, "Three-phase solid oxide fuel cell anode microstructure realization using two-point correlation functions," *Acta materialia,* vol. 59, no. 1, pp. 30-43, 2011.

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

[11] H. Kumar, C. Briant, and W. Curtin, "Using microstructure reconstruction to model mechanical behavior in complex microstructures," *Mechanics of Materials,* vol. 38, no. 8, pp. 818-832, 2006.

[12] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics,* vol. 9, no. 1, pp. 62-66, 1979.

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

[14] M. S. Greene, Y. Liu, W. Chen, and W. K. Liu, "Computational uncertainty analysis in multiresolution materials via stochastic constitutive theory," *Computer Methods in Applied Mechanics and Engineering,* vol. 200, no. 1, pp. 309-325, 2011.

[15] B. HKDH, "Neural networks in materials science," *ISIJ international,* vol. 39, no. 10, pp. 966-979, 1999.

[16] V. Sundararaghavan and N. Zabaras, "Classification and reconstruction of three-dimensional microstructures using support vector machines," *Computational Materials Science,* vol. 32, no. 2, pp. 223-239, 2005.

[17] R. Bostanabad, Zhang, Y., Li, X., Kearney, T., Brinson, L. C., Apley, , & Daniel W., L., Wing K., and Chen, W., "Computational Microstructure Characterization and Reconstruction: Review of the State-of-the-art Techniques," *Progress in Materials Science,* accepted (https://doi.org/10.1016/j.pmatsci.2018.01.005) 2018

[18] Y. Liu, M. S. Greene, W. Chen, D. A. Dikin, and W. K. Liu, "Computational microstructure characterization and reconstruction for stochastic multiscale material design," *Computer-Aided Design,* vol. 45, no. 1, pp. 65-76, 2013.

[19] C. Yeong and S. Torquato, "Reconstructing random media. II. Three-dimensional media from two-dimensional cuts," *Physical Review E,* vol. 58, no. 1, p. 224, 1998.

[20] C. Yeong and S. Torquato, "Reconstructing random media," *Physical Review E,* vol. 57, no. 1, p. 495, 1998.

[21] 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.

[22] 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.

[23] R. Bostanabad, A. T. Bui, W. Xie, D. W. Apley, and W. Chen, "Stochastic microstructure characterization and reconstruction via supervised learning," *Acta Materialia,* vol. 103, pp. 89-102, 2016.

[24] R. Cang, Y. Xu, S. Chen, Y. Liu, Y. Jiao, and M. Y. Ren, "Microstructure Representation and Reconstruction of Heterogeneous Materials via Deep Belief Network for Computational Material Design," *Journal of Mechanical Design,* vol. 139, no. 7, p. 071404, 2017.

[25] S. C. Yu *et al.*, "Characterization and Design of Functional Quasi-Random Nanostructured Materials Using Spectral Density Function," *Journal of Mechanical Design, 139(7), 071401. https://doi.org/10.1115/1.4036582,* vol. 139, no. July, pp. 135-145, 2016.

[26] S. Torquato, "Disordered hyperuniform heterogeneous materials," *Journal of Physics: Condensed Matter,* vol. 28, no. 41, p. 414012, 2016.

[27] 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.

[28] O. U. Uche, F. H. Stillinger, and S. Torquato, "Constraints on collective density variables: Two dimensions," *Physical Review E,* vol. 70, no. 4, p. 046122, 2004.

[29] O. U. Uche, S. Torquato, and F. H. Stillinger, "Collective coordinate control of density distributions," *Physical Review E,* vol. 74, no. 3, p. 031104, 2006.

[30] R. D. Batten, F. H. Stillinger, and S. Torquato, "Classical disordered ground states: Super-ideal gases and stealth and equi-luminous materials," *Journal of Applied Physics,* vol. 104, no. 3, p. 033504, 2008.

[31] M. Florescu, S. Torquato, and P. J. Steinhardt, "Designer disordered materials with large, complete photonic band gaps," *Proceedings of the National Academy of Sciences,* vol. 106, no. 49, pp. 20658-20663, 2009.

[32] R. Bostanabad, W. Chen, and D. Apley, "Characterization and reconstruction of 3D stochastic microstructures via supervised learning," *Journal of microscopy,* vol. 264, no. 3, pp. 282-297, 2016.

[33] J. Fu, S. Cui, S. Cen, and C. Li, "Statistical characterization and reconstruction of heterogeneous microstructures using deep neural network," *Computer Methods in Applied Mechanics and Engineering,* vol. 373, p. 113516, 2021.

[34] R. Bostanabad, "Reconstruction of 3d microstructures from 2d images via transfer learning," *Computer-Aided Design,* vol. 128, p. 102906, 2020.

[35] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, "A transfer learning approach for microstructure reconstruction and structure-property predictions," *Scientific reports,* vol. 8, 2018.

[36] X. Liu and V. Shapiro, "Random heterogeneous materials via texture synthesis," *Computational Materials Science,* vol. 99, pp. 177-189, 2015.

[37] V. Sundararaghavan, "Reconstruction of three-dimensional anisotropic microstructures from two-dimensional micrographs imaged on orthogonal planes," *Integrating Materials and Manufacturing Innovation,* vol. 3, no. 1, p. 19, 2014.

[38] R. Cang, H. Li, H. Yao, Y. Jiao, and Y. Ren, "Improving direct physical properties prediction of heterogeneous materials from imaging data via convolutional neural network and a morphology-aware generative model," *Computational Materials Science,* vol. 150, pp. 212-221, 2018.

[39] Z. Yang, X. Li, L. Catherine Brinson, A. N. Choudhary, W. Chen, and A. Agrawal, "Microstructural Materials Design Via Deep Adversarial Learning Methodology," *Journal of Mechanical Design,* vol. 140, no. 11, pp. 111416-111416-10, 2018.

[40] A. Iyer, B. Dey, A. Dasgupta, W. Chen, and A. Chakraborty, "A Conditional Generative Model for Predicting Material Microstructures from Processing Methods," *arXiv preprint arXiv:1910.02133,* 2019.

[41] J. W. Lee, N. H. Goo, W. B. Park, M. Pyo, and K. S. Sohn, "Virtual microstructure design for steels using generative adversarial networks," *Engineering Reports,* vol. 3, no. 1, p. e12274, 2021.

[42] B. Dong *et al.*, "Optical response of a disordered bicontinuous macroporous structure in the longhorn beetle Sphingnotus mirabilis," *Physical Review E,* vol. 84, no. 1, p. 011915, 2011.

[43] W.-K. Lee, C. J. Engel, M. D. Huntington, J. Hu, and T. W. Odom, "Controlled three-dimensional hierarchical structuring by memory-based, sequential wrinkling," *Nano letters,* vol. 15, no. 8, pp. 5624-5629, 2015.

[44] C. Chatfield, *The analysis of time series: an introduction*. CRC press, 2016.

[45] G. J. Hedley *et al.*, "Determining the optimum morphology in high-performance polymer-fullerene organic photovoltaic cells," *Nature Communications,* vol. 4, p. 2867, 2013.

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

[47] J. W. Cahn, "Phase separation by spinodal decomposition in isotropic systems," *The Journal of Chemical Physics,* vol. 42, no. 1, pp. 93-99, 1965.

[48] Y. Zhang, S. Tao, W. Chen, and D. W. Apley, "A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors," *arXiv preprint arXiv:1806.07504,* 2018.

[49] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*, 2003, pp. 63-71: Springer.

[50] R. Bostanabad, T. Kearney, S. Tao, D. W. Apley, and W. Chen, "Leveraging the nugget parameter for efficient Gaussian process modeling," *International Journal for Numerical Methods in Engineering,* vol. 114, no. 5, pp. 501-516, 2018.

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

[52] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[53] H. J. Kushner, "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise," *Journal of Basic Engineering,* vol. 86, no. 1, pp. 97-106, 1964.

[54] 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.

[55] 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.

[56] N. Espinosa, M. Hösel, D. Angmo, and F. C. Krebs, "Solar cells with one-day energy payback for the factories of the future," *Energy & Environmental Science,* vol. 5, no. 1, pp. 5117-5132, 2012.

[57] Y. Li, G. Xu, C. Cui, and Y. Li, "Flexible and semitransparent organic solar cells," *Advanced Energy Materials,* vol. 8, no. 7, p. 1701791, 2018.

[58] E. K. Lee, M. Y. Lee, C. H. Park, H. R. Lee, and J. H. Oh, "Toward environmentally robust organic electronics: approaches and applications," *Advanced Materials,* vol. 29, no. 44, p. 1703638, 2017.

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

[60] M. T. Dang, L. Hirsch, and G. Wantz, "P3HT: PCBM, best seller in polymer photovoltaic research," *Advanced Materials,* vol. 23, no. 31, pp. 3597-3602, 2011.

[61] G. Grancini, D. Polli, D. Fazzi, J. Cabanillas-Gonzalez, G. Cerullo, and G. Lanzani, "Transient absorption imaging of P3HT: PCBM photovoltaic blend: Evidence for interfacial charge transfer state," *The Journal of Physical Chemistry Letters,* vol. 2, no. 9, pp. 1099-1105, 2011.

[62] J. Song, M. Zhang, M. Yuan, Y. Qian, Y. Sun, and F. Liu, "Morphology Characterization of Bulk Heterojunction Solar Cells," *Small Methods,* vol. 2, no. 3, p. 1700229, 2018.

[63] L. Lu, T. Zheng, Q. Wu, A. M. Schneider, D. Zhao, and L. Yu, "Recent advances in bulk heterojunction polymer solar cells," *Chemical reviews,* vol. 115, no. 23, pp. 12666-12731, 2015.

[64] D. T. Fullwood, S. R. Niezgoda, B. L. Adams, and S. R. Kalidindi, "Microstructure sensitive design for performance optimization," *Progress in Materials Science,* vol. 55, no. 6, pp. 477-562, 2010.

[65] J. H. Panchal, S. R. Kalidindi, and D. L. McDowell, "Key computational modeling issues in integrated computational materials engineering," *Computer-Aided Design,* vol. 45, no. 1, pp. 4-25, 2013.

[66] 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.

[67] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments," *Statistical science,* pp. 409-423, 1989.

[68] J. R. Tumbleston, D.-H. Ko, E. T. Samulski, and R. Lopez, "Nonideal parasitic resistance effects in bulk heterojunction organic solar cells," *Journal of Applied Physics,* vol. 108, no. 8, p. 084514, 2010.

[69] O. V. Mikhnenko, H. Azimi, M. Scharber, M. Morana, P. W. M. Blom, and M. A. Loi, "Exciton diffusion length in narrow bandgap polymers," *Energy & Environmental Science,* vol. 5, no. 5, p. 6960, 2012.

[70] V. D. Mihailetchi, H. X. Xie, B. De Boer, L. J. A. Koster, and P. W. M. Blom, "Charge Transport and Photocurrent Generation in Poly(3‐hexylthiophene): Methanofullerene Bulk‐Heterojunction Solar Cells," *Advanced Functional Materials,* vol. 16, no. 5, pp. 699-708, 2006.

[71] S. H. Park *et al.*, "Bulk heterojunction solar cells with internal quantum efficiency approaching 100%," *Nature photonics,* vol. 3, no. 5, p. 297, 2009.

[72] M. Reyes-Reyes, K. Kim, and D. L. Carroll, "High-efficiency photovoltaic devices based on annealed poly(3-hexylthiophene) and 1-(3-methoxycarbonyl)-propyl-1- phenyl-(6,6)C61 blends," *Applied Physics Letters,* vol. 87, no. 8, p. 083506, 2005.

[73] P. Berger and M. Kim, "Polymer solar cells: P3HT: PCBM and beyond," *Journal of Renewable and Sustainable Energy,* vol. 10, no. 1, 2018.

[74] R. Jin, W. Chen, and A. Sudjianto, "An efficient algorithm for constructing optimal design of computer experiments," *Journal of Statistical Planning and Inference,* vol. 134, no. 1, pp. 268-287, 2005.

[75] L. Tan and J. Jiang, "Chapter 3 - Digital Signals and Systems," in *Digital Signal Processing (Second Edition)*, L. Tan and J. Jiang, Eds. Boston: Academic Press, 2013, pp. 57-85.

[76] J. A. Gubner, *Probability and random processes for electrical and computer engineers*. Cambridge University Press, 2006.

[77] R. Ansari and L. Valbonesi, "1 - Signals and Systems," in *The Electrical Engineering Handbook*, W.-K. Chen, Ed. Burlington: Academic Press, 2005, pp. 813-837.

[78] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," *npj Computational Materials,* vol. 5, no. 1, p. 21, 2019.

[79] A. Iyer *et al.*, "Data-Centric Mixed-Variable Bayesian Optimization For Materials Design," *arXiv preprint arXiv:1907.02577,* 2019.

[80] 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.

[81] 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.

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

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] G. Chen *et al.*, "Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges," *Polymers,* vol. 12, no. 1, p. 163, 2020.

[94] Y. Z. A. Iyer, A. Prasad, S. Tao, Y. Wang, L. Schadler, L.C. Brinson, W. Chen, "Data Centric Mixed Variable Bayesian Optimization for Materials Design," *ASME International Design Engineering Technical Conference,* 2019.

[95] 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.

[96] 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.

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

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

[99] 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.

[100] 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.

[101] 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.

[102] Y. Wang *et al.*, "Mining structure–property relationships in polymer nanocomposites using data driven finite element analysis and multi-task convolutional neural networks," *Molecular Systems Design & Engineering,* vol. 5, no. 5, pp. 962-975, 2020.

[103] 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.

[104] 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.

[105] 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.

[106] L. Schadler *et al.*, "A perspective on the data-driven design of polymer nanodielectrics," *Journal of Physics D: Applied Physics,* vol. 53, no. 33, p. 333001, 2020.

[107] 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.

[108] 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.

[109] 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.

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

[111] 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.

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

[113] 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.

[114] A. S. Prasad *et al.*, "Investigating the effect of surface modification on the dispersion process of polymer nanocomposites," *Nanocomposites,* vol. 6, no. 3, pp. 111-124, 2020.

[115] 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.

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

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

[118] M. Imada, A. Fujimori, and Y. Tokura, "Metal-insulator transitions," *Reviews of modern physics,* vol. 70, no. 4, p. 1039, 1998.

[119] N. Shukla *et al.*, "A steep-slope transistor based on abrupt electronic phase transition," *Nature communications,* vol. 6, no. 1, pp. 1-6, 2015.

[120] Y. Zhou and S. Ramanathan, "Mott memory and neuromorphic devices," *Proceedings of the IEEE,* vol. 103, no. 8, pp. 1289-1310, 2015.

[121] Z. Yang, C. Ko, and S. Ramanathan, "Oxide electronics utilizing ultrafast metal-insulator transitions," *Annual Review of Materials Research,* vol. 41, pp. 337-367, 2011.

[122] W. Zhang, J. Liu, and T.-C. Wei, "Machine learning of phase transitions in the percolation and X Y models," *Physical Review E,* vol. 99, no. 3, p. 032142, 2019.

[123] M. Coll *et al.*, "Towards oxide electronics: a roadmap," *Applied surface science,* vol. 482, pp. 1-93, 2019.

[124] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature,* vol. 559, no. 7715, pp. 547-555, 2018.

[125] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Physical review letters,* vol. 120, no. 14, p. 145301, 2018.

[126] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials,* vol. 5, no. 1, pp. 1-36, 2019.

[127] J. Noh *et al.*, "Inverse design of solid-state materials via a continuous representation," *Matter,* vol. 1, no. 5, pp. 1370-1384, 2019.

[128] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization," *Physical review letters,* vol. 115, no. 20, p. 205901, 2015.

[129] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, "High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates," *Integrating Materials and Manufacturing Innovation,* vol. 6, no. 3, pp. 207-217, 2017.

[130] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, "Multi-objective Optimization for Materials Discovery via Adaptive Design," *Scientific reports,* vol. 8, no. 1, p. 3738, 2018.

[131] L. Cario, C. Vaju, B. Corraze, V. Guiot, and E. Janod, "Electric-Field-Induced Resistive Switching in a Family of Mott Insulators: Towards a New Class of RRAM Memories," *Advanced Materials,* vol. 22, no. 45, pp. 5193-5197, 2010.

[132] A. V. Powell, A. McDowall, I. Szkoda, K. S. Knight, B. J. Kennedy, and T. Vogt, "Cation Substitution in Defect Thiospinels: Structural and Magnetic Properties of GaV4-x Mo x S8 (0⩽ x⩽ 4)," *Chemistry of Materials,* vol. 19, no. 20, pp. 5035-5044, 2007.

[133] A. Camjayi *et al.*, "First-order insulator-to-metal Mott transition in the paramagnetic 3D system GaTa 4 Se 8," *Physical review letters,* vol. 113, no. 8, p. 086404, 2014.

[134] I. Kézsmárki *et al.*, "Néel-type skyrmion lattice with confined orientation in the polar magnetic semiconductor GaV 4 S 8," *Nature materials,* vol. 14, no. 11, p. 1116, 2015.

[135] Z. Wang *et al.*, "Polar Dynamics at the Jahn-Teller Transition in Ferroelectric Ga V 4 S 8," *Physical review letters,* vol. 115, no. 20, p. 207601, 2015.

[136] E. Dorolti *et al.*, "Half-metallic ferromagnetism and large negative magnetoresistance in the new lacunar spinel GaTi3VS8," *Journal of the American Chemical Society,* vol. 132, no. 16, pp. 5704-5710, 2010.

[137] R. Pocha, D. Johrendt, and R. Pöttgen, "Electronic and structural instabilities in GaV4S8 and GaMo4S8," *Chemistry of materials,* vol. 12, no. 10, pp. 2882-2887, 2000.

[138] M. Sieberer, S. Turnovszky, J. Redinger, and P. Mohn, "Importance of cluster distortions in the tetrahedral cluster compounds Ga M 4 X 8 (M= Mo, V, Nb, Ta; X= S, Se): Ab initio investigations," *Physical Review B,* vol. 76, no. 21, p. 214106, 2007.

[139] D. Bichler and D. Johrendt, "Tuning of Metal− Metal Bonding and Magnetism via the Electron Count in Ga x V4-y Cr y S8," *Chemistry of materials,* vol. 19, no. 17, pp. 4316-4321, 2007.

[140] C. J. Bartel *et al.*, "Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry," *Nature communications,* vol. 9, no. 1, pp. 1-10, 2018.

[141] J. K. Burdett, B. A. Coddens, and G. V. Kulkarni, "Band gap and stability of solids," *Inorganic chemistry,* vol. 27, no. 18, pp. 3259-3261, 1988.

[142] M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics,* vol. 21, no. 2, pp. 239-245, 1979.

[143] B. Corraze *et al.*, "Electric field induced avalanche breakdown and non-volatile resistive switching in the Mott Insulators AM 4 Q 8," *The European Physical Journal Special Topics,* vol. 222, no. 5, pp. 1046-1056, 2013.

[144] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, "Thermodynamic limit for synthesis of metastable inorganic materials," *Science advances,* vol. 4, no. 4, p. eaaq0148, 2018.

[145] S. V. Streltsov and D. I. Khomskii, "Covalent bonds against magnetism in transition metal compounds," *Proceedings of the National Academy of Sciences,* vol. 113, no. 38, pp. 10491-10496, 2016.

[146] C. Vaju *et al.*, "Electric-Pulse-driven Electronic Phase Separation, Insulator–Metal Transition, and Possible Superconductivity in a Mott Insulator," *Advanced Materials,* vol. 20, no. 14, pp. 2760-2765, 2008.

[147] D. M. Juraschek, M. Fechner, and N. A. Spaldin, "Ultrafast structure switching through nonlinear phononics," *Physical review letters,* vol. 118, no. 5, p. 054101, 2017.

[148] A. Zunger, "Inverse design in search of materials with target functionalities," *Nature Reviews Chemistry,* vol. 2, no. 4, pp. 1-16, 2018.

[149] H. C. Herbol, W. Hu, P. Frazier, P. Clancy, and M. Poloczek, "Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization," *npj Computational Materials,* vol. 4, no. 1, p. 51, 2018.

[150] R. Yuan *et al.*, "Accelerated Search for BaTiO3-Based Ceramics with Large Energy Storage at Low Fields Using Machine Learning and Experimental Design," *Advanced Science*.

[151] B. J. Shields *et al.*, "Bayesian reaction optimization as a tool for chemical synthesis," *Nature,* vol. 590, no. 7844, pp. 89-96, 2021.

[152] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, and K. Tsuda, "COMBO: an efficient Bayesian optimization library for materials science," *Materials discovery,* vol. 4, pp. 18-21, 2016.

[153] F. Häse, L. M. Roch, and A. Aspuru-Guzik, "Gryffin: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry," *arXiv preprint arXiv:2003.12127,* 2020.

[154] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton, "High-throughput computational screening of perovskites for thermochemical water splitting applications," *Chemistry of Materials,* vol. 28, no. 16, pp. 5621-5634, 2016.

[155] J. Li *et al.*, "Highly Anisotropic P3HT Film Fabricated via Epitaxy on an Oriented Polyethylene Film and Solvent Vapor Treatment," *Langmuir,* vol. 35, no. 24, pp. 7841-7847, 2019/06/18 2019.

[156] R. Dulal *et al.*, "Elongated Nano Domains and Molecular Intermixing induced Doping in Organic Photovoltaic Active Layers with Electric Field Treatment," *ACS Applied Polymer Materials*.

[157] J. Munshi *et al.*, "Composition and processing dependent miscibility of P3HT and PCBM in organic solar cells by coarse-grained molecular simulations," *Computational Materials Science,* vol. 155, pp. 112-115, 2018.

[158] J. Munshi, R. Dulal, T. Chien, W. Chen, and G. Balasubramanian, "Solution processing dependent bulk heterojunction nanomorphology of P3HT: PCBM thin films," *ACS applied materials & interfaces*.

[159] J. Munshi *et al.*, "Effect of polydispersity on the bulk-heterojunction morphology of P3HT: PCBM solar cells," *Journal of Polymer Science Part B: Polymer Physics,* vol. 57, no. 14, pp. 895-903, 2019.

[160] A. van Beek *et al.*, "Scalable Adaptive Batch Sampling in Simulation-Based Design With Heteroscedastic Noise," *Journal of Mechanical Design,* vol. 143, no. 3, p. 031709, 2021.

[161] A. Van Beek *et al.*, "Scalable Objective-Driven Batch Sampling in Simulation-Based Design for Models With Heteroscedastic Noise," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2020, vol. 84010, p. V11BT11A049: American Society of Mechanical Engineers.

[162] B. L. DeCost, M. D. Hecht, T. Francis, B. A. Webler, Y. N. Picard, and E. A. Holm, "UHCSDB: UltraHigh carbon steel micrograph database," *Integrating Materials and Manufacturing Innovation,* vol. 6, no. 2, pp. 197-205, 2017.

[163] X. Zhao, A. Iyer, P. Promoppatum, and S.-C. Yao, "Numerical modeling of the thermal behavior and residual stress in the direct metal laser sintering process of titanium alloy products," *Additive Manufacturing,* vol. 14, pp. 126-136, 2017.

[164] P. Promoppatum, R. Onler, and S.-C. Yao, "Numerical and experimental investigations of micro and macro characteristics of direct metal laser sintered Ti-6Al-4V products," *Journal of Materials Processing Technology,* vol. 240, pp. 262-273, 2017.

[165] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[166] Z. Wang *et al.*, "Effect of high aspect ratio filler on dielectric properties of polymer composites: a study on barium titanate fibers and graphene platelets," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 19, no. 3, pp. 960-967, 2012.

[167] Z. Wang, J. K. Nelson, H. Hillborg, S. Zhao, and L. S. Schadler, "Dielectric constant and breakdown strength of polymer composites with high aspect ratio fillers studied by finite element models," *Composites science and technology,* vol. 76, pp. 29-36, 2013.

[168] J. J. Mortensen, K. Kaasbjerg, S. L. Frederiksen, J. K. Nørskov, J. P. Sethna, and K. W. Jacobsen, "Bayesian error estimation in density-functional theory," *Physical review letters,* vol. 95, no. 21, p. 216401, 2005.

[169] J. Wellendorff *et al.*, "Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation," *Physical Review B,* vol. 85, no. 23, p. 235149, 2012.

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

[171] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical review B,* vol. 54, no. 16, p. 11169, 1996.

[172] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," *Physical review b,* vol. 59, no. 3, p. 1758, 1999.

[173] P. E. Blöchl, "Projector augmented-wave method," *Physical review B,* vol. 50, no. 24, p. 17953, 1994.

[174] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical review letters,* vol. 77, no. 18, p. 3865, 1996.

[175] Y. Wang, D. Puggioni, and J. M. Rondinelli, "Assessing exchange-correlation functional performance in the chalcogenide lacunar spinels GaM $ _4 $ Q $ _8 $(M= Mo, V, Nb, Ta; Q= S, Se)," *arXiv preprint arXiv:1905.09170,* 2019.

[176] H.-S. Kim, J. Im, M. J. Han, and H. Jin, "Spin-orbital entangled molecular j eff states in lacunar spinel compounds," *Nature communications,* vol. 5, no. 1, pp. 1-7, 2014.

[177] A. Rastogi *et al.*, "Itinerant electron magnetism in the Mo 4 tetrahedral cluster compounds GaMo 4 S 8, GaMo 4 Se 8, and GaMo 4 Se 4 Te 4," *Journal of low temperature physics,* vol. 52, no. 5, pp. 539-557, 1983.

[178] P. E. Blöchl, O. Jepsen, and O. K. Andersen, "Improved tetrahedron method for Brillouin-zone integrations," *Physical Review B,* vol. 49, no. 23, p. 16223, 1994.

[179] R. Resta, "Macroscopic polarization in crystalline dielectrics: the geometric phase approach," *Reviews of modern physics,* vol. 66, no. 3, p. 899, 1994.

[180] A. Togo and I. Tanaka, "First principles phonon calculations in materials science," *Scripta Materialia,* vol. 108, pp. 1-5, 2015.

[181] S. Kirklin, B. Meredig, and C. Wolverton, "High-throughput computational screening of new Li-ion battery anode materials," *Advanced Energy Materials,* vol. 3, no. 2, pp. 252-262, 2013.

[182] S. Kirklin *et al.*, "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies," *npj Computational Materials,* vol. 1, no. 1, pp. 1-15, 2015.

[183] G. K. Madsen, J. Carrete, and M. J. Verstraete, "BoltzTraP2, a program for interpolating band structures and calculating semi-classical transport coefficients," *Computer Physics Communications,* vol. 231, pp. 140-145, 2018.

[184] A. Iyer *et al.*, "Data centric nanocomposites design via mixed-variable Bayesian optimization," *Molecular Systems Design & Engineering,* vol. 5, no. 8, pp. 1376-1390, 2020.

[185] Y. Wang, A. Iyer, W. Chen, and J. M. Rondinelli, "Featureless adaptive optimization accelerates functional electronic materials design," *Applied Physics Reviews,* vol. 7, no. 4, p. 041403, 2020.

9 Appendix

9.1 One-Click Microstructure Characterization and Reconstruction via NanoMine

An important component in the recent push towards data-driven design of materials has been the emergence of open-source material databases [3-7] providing quick access to materials' data in a machine readable format. NanoMine [3, 4], a nanocomposite material database with in-built data curation, exploration and analysis capabilities, exemplifies this approach in the field of polymer nanocomposites. It captures the physical properties reported in the literature and from individual research labs including microstructure, processing conditions, and material properties. Ontology-enabled knowledge graph framework helps NanoMine establish relationship between those properties. A collection of module tools for microstructure characterization & reconstruction and simulation software to model bulk nanocomposite material response augments knowledge generated by experimental data. Integrating these different sources of knowledge is critical for establishing PSP relationships and subsequently material design.

Commonly used MCR tools have been incorporated in NanoMine to provide parsimonious microstructures analysis workflow for researchers as shown in Figure 9-1. As mentioned in Chapter 2.1, microstructure binarization is a precursor to MCR and has thus been included in NanoMine. We provide two popular binarization tools namely, Otsu and Niblack's Method and three microstructure characterization and reconstruction techniques applicable for two-phase nanocomposites.

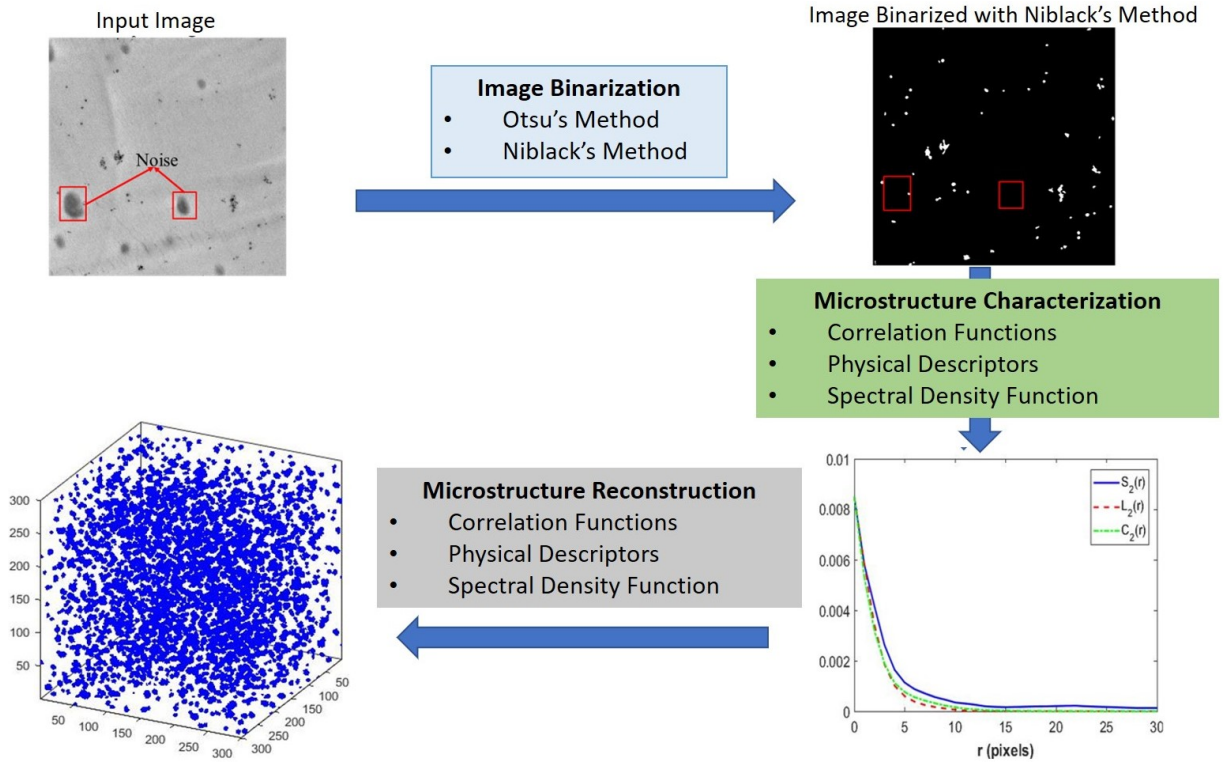


Figure 9-1: Summary of microstructure binarization, characterization and reconstruction tools offered by NanoMine.

Each tool is accompanied by detailed instruction to users regarding how to interact with the tool as well as recommendations for how to select tools best suited for their microstructure. Some user-friendly features that make these tools attractive for researchers are summarized below:

- All computations are performed on the NanoMine server. Thus, users can have full access to webtools via their web browser and do not need to install any software packages.
- All tools support commonly used image file formats such as Joint Photographic Exchange Group (JPEG) / Portable Network Graphics (PNG) / Tagged Image File Format (TIFF) as well as MATLAB's native .mat file format.
- If the user wants to analyze a batch of images, NanoMine provides the option of uploading a ZIP file containing several images. The images must belong to one of the supported file

formats mentioned above. NanoMine tools process each image separately and then compute pertinent statistics over the entire batch. It is essential that all images have the same scale and resolution for accurate computation of batch statistics. All data (statistics for each image as well as aggregated statistics for the batch) is returned to the user for in comma-separated values (CSV) file format for inference and further analysis.

- Every tool request made by the user is assigned a unique Job ID, which can be used to retrieve results. The Job ID is a 22 character alphanumeric code generated upon receiving the job request and is used to store all files created during tool execution.
- Since some tools are computationally intensive and require substantial time for completion, NanoMine provides E-mail alerts to users upon completion of their requests. This feature alleviates the need for users to stay connected with NanoMine server during processing of their requests. An E-mail containing the status of request (successful completion/failure), the Job ID and a hyperlink to results page is sent to address provided by user during registration on NanoMine website. The results page provides a summary of results obtained. For example, Figure 9-2 shows the result page for microstructure characterization using correlation webtool. The uploaded image & its correlation are shown on screen and also available for download for further analysis.
- The outputs from each tool are returned to users in a machine-readable format, helpful for further analysis.

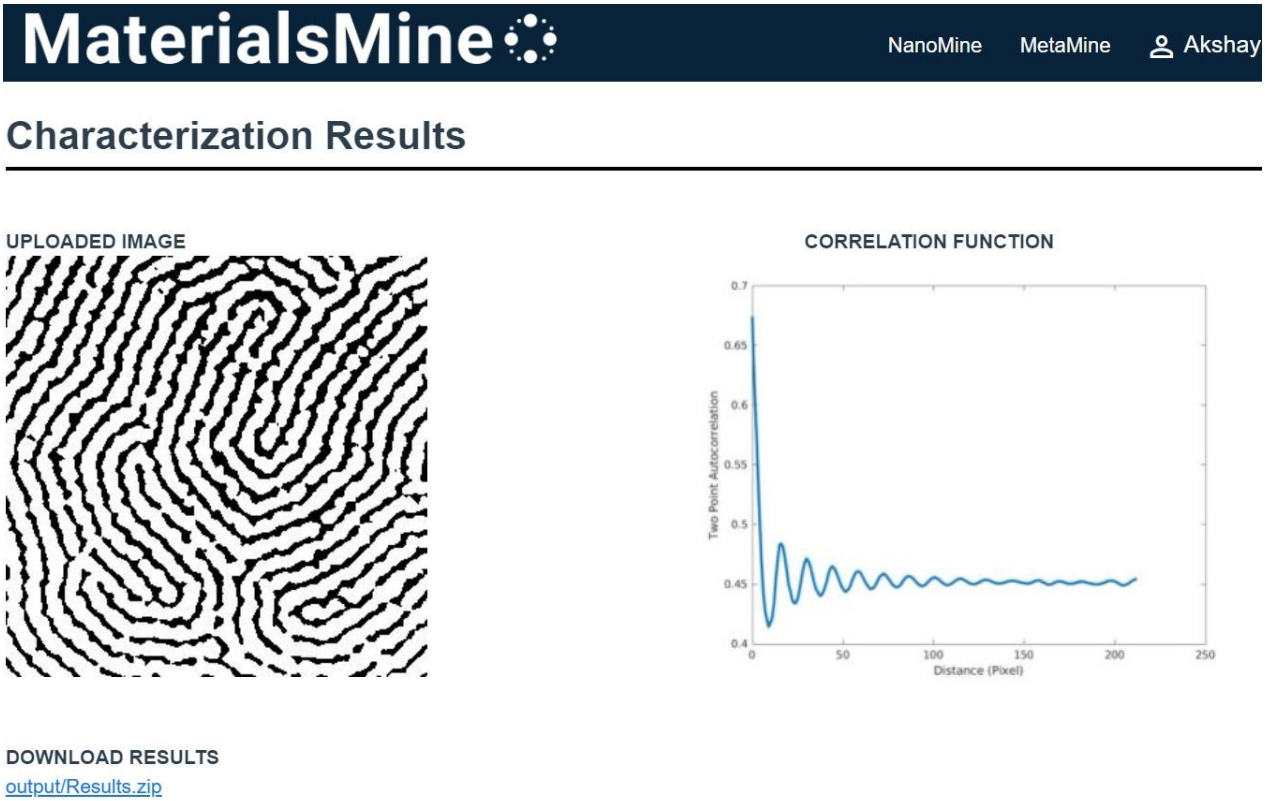


Figure 9-2: Snapshot of result obtained from Correlation function characterization tool in NanoMine

9.2 Interphase Calibration for Polymer Nanocomposites

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 ($\Delta\varepsilon_i$)

$$\varepsilon'(\omega) = \varepsilon_\infty + \sum_{i=1}^{n} \frac{\Delta\varepsilon_i}{1 + (\omega\tau_i)^2}, \tag{9-1}$$

$$\varepsilon''(\omega) = \sum_{i=1}^{n} \frac{\Delta\varepsilon_i \omega\tau_i}{1 + (\omega\tau_i)^2}, \tag{9-2}$$

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.

$$\varepsilon'_{int}(\omega) = \varepsilon_\infty + C + M_\alpha \sum_{\tau_i>\tau_0} \frac{\Delta\varepsilon_i}{1 + (\omega S_\alpha\tau_i)^2} + M_\beta \sum_{\tau_i<\tau_0} \frac{\Delta\varepsilon_i}{1 + (\omega S_\beta\tau_i)^2}, \tag{9-3}$$

$$\varepsilon''_{int}(\omega) = M_\alpha \sum_{\tau_i>\tau_0} \frac{\Delta\varepsilon_i \omega S_\alpha\tau_i}{1 + (\omega S_\alpha\tau_i)^2} + M_\beta \sum_{\tau_i<\tau_0} \frac{\Delta\varepsilon_i \omega S_\beta\tau_i}{1 + (\omega S_\beta\tau_i)^2}, \tag{9-4}$$

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 our 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.

$$\varepsilon'_{int}(\omega = 60Hz) = SF_{real} * \varepsilon'(\omega = 60hz), \tag{9-5}$$

$$\varepsilon''_{int}(\omega = 60Hz) = SF_{imag} * \varepsilon''(\omega = 60hz), \tag{9-6}$$

Calibration of these scale factors (Module 2, Chapter 4) 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 [80, 115] or

using black-box optimization methods, for instance, adaptive sampling using Bayesian approach [170], 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].

9.3 Density Functional Calculation Details for Lacunar Spinels

We perform DFT simulations as implemented in the Vienna *Ab initio* Simulation Package (VASP) [171, 172]. The projector augmented-wave (PAW) potentials [173] are used for all elements in our calculations with the following valence electron configurations: Al ($3s^2 3p^1$), Ga ($3d^{10} 4s^2 4p^1$), In ($4d^{10} 5s^2 5p^1$), V ($3s^2 3p^6 3d^4 4s^1$), Nb ($4s^2 4p^6 4d^4 5s^1$), Ta ($5p^6 5d^4 6s^1$), Cr ($3s^2 3p^6 3d^5 4s^1$), Mo ($4s^2 4p^6 4d^5 5s^1$), W($5s^2 5p^6 5d^5 6s^1$), S ($3s^2 3p^4$), Se ($4s^2 4p^4$), and Te ($5s^2 5p^4$). We use exchange correlation potentials (V_{xc}) as implemented by Perdew-Burke-Ernzerhof (PBE) [174]. The effect of on-site Coulomb interactions (PBE+U) is considered with a U value of 2.0eV for all 6 transition metals. Previous studies have shown that such settings could nicely capture the complex electronic structures within the lacunar spinel family [175, 176]. Numerous spin configurations are evaluated to ensure the global ground state is achieved and that

those states are consistent with available experimental magnetic data [177]. Spin-orbit interactions (SOI) are not considered in our calculations. Although it has been shown that SOI leads to interesting molecular j_{eff} states [176], this order does not strongly affect the size of the ground state electronic band gaps, even $5d$ transition metals lacunar spinels [175]. A Γ-centered 6 × 6 × 6 k-point mesh with a 500eV kinetic energy cutoff is used. We employ Gaussian smearing with a small 0.05eV width. For density-of-state calculations, we use the tetrahedron method with Blöchl corrections [178]. Electric polarizations along the [111] direction are simulated using the Berry phase method [179].

The crystal structures of the existing lacunar spinels are obtained from our previous DFT studies[2], structures of new compositions are obtained by replacing the elements on the corresponding crystallographic sites from existing structures. We perform full lattice relaxations until the residual forces on each individual atom are less than 1.0meVÅ^{-1}. The DFT relaxed crystal structures of the Pareto front compositions are available on GitHub[3]. We initialize the relaxation with various magnetic moment configurations, the converged configuration with the lowest energy is reported as the DFT ground state. Zone center (k=0) phonon frequencies and eigendisplacements are obtained using the frozen-phonon method with pre and post-processing performed with the Phonopy package [180]. The decomposition pathways are automatically generated using Grand Canonical Linear Programming [181] from the Open Quantum Materials Database [182].

Resistivity simulations are performed using electronic structures computed from VASP as previously described, but with an increased 24 × 24 × 24 k-point mesh and the BoltzTrap2 package

[2] GitHub Repository

[3] GitHub Repository

[183]. We also assume that all M^a sites have the same orientation within the crystal. In order to validate this model, we simulated a 2 × 2 × 2 supercell of $InNbMo_3Se_8$ with one Nb atom oriented in a different direction from the other seven. We find that the ground state E_g as well as ΔH_d exhibit negligible changes from the homogeneous description. We also compared the change in properties with the anti-ferromagnetic spin configuration using a doubled simulation cell with the ferromagnetic ground state. As before, we find there are no significant changes in the aforementioned properties. These results are reasonable because the local structure of the transition metal cluster dictates the low-energy band structure near the Fermi level.

9.4 Cover art for Data Centric Nanocomposite Design article

The article describing data-centric polymer nanocomposite design using Bayesian Optimization was highlighted on the cover of Molecular System Design and Engineering journal [184]. A copy of the published cover art is shown in Figure 9-3.

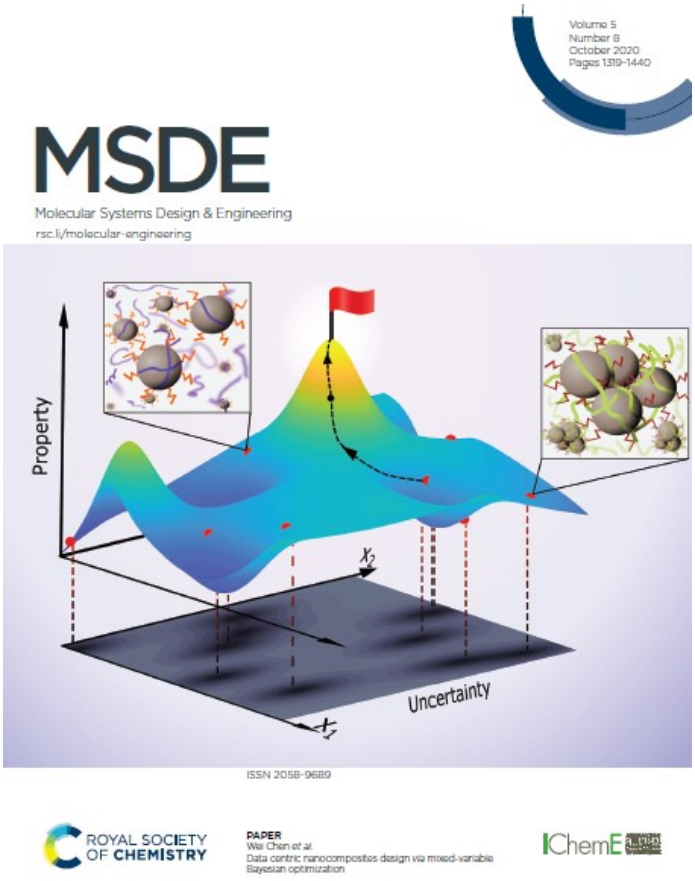


Figure 9-3: Copy of Molecular System Design and Engineering journal cover highlighting our article

9.5 Metal Insulator Transitions design article in the news

The article describing features optimization of Metal Insulator Transition compounds was highlighted on the cover of Applied Physics Review [185] journal (Figure 9-4a) and covered by

Northwestern University's McCormick School of Engineering in their weekly newsletter (Figure 9-4b).

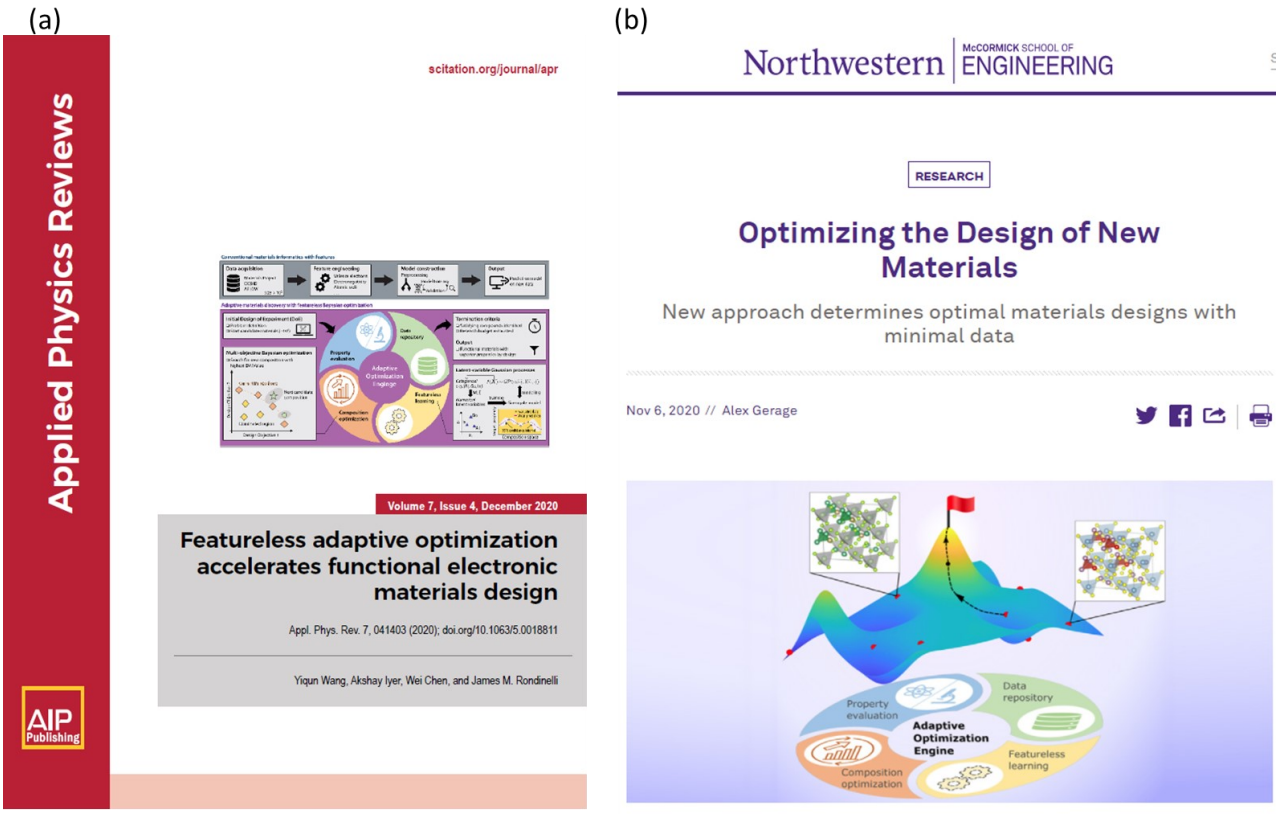


Figure 9-4: (a) Copy of Applied Physics Review journal cover (b) A screenshot of news article published by McCormick School of Engineering.