NORTHWESTERN UNIVERSITY

Combining Finite Element with Data Analytical Approaches for Structure-Property Modeling in Polymer Nanocomposites

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Mechanical Engineering

By

Yixing Wang

EVANSTON, ILLINOIS

September 2019

ABSTRACT

Polymer nanocomposites have attracted great interest in recent years because of their potential as tailored materials with enhanced properties. Recent experiments have shown that polymer nanocomposites are able to achieve significant improvement in dielectrical, thermal, mechanical and other physical properties compared with their parent polymer systems. More importantly, these outstanding properties can be achieved at low filler loadings such that the polymer system does not sacrifice the advantages of easy processability.

Despite the excellent performance of advanced materials including polymer nanocomposites, the time frame of development to application of those materials in industry is long, typically over 10 years. Therefore, a top priority of researchers and engineers is to reduce the time required to bring advanced materials from laborites to the market. The Material Genome Initiative (MGI), proposed by the White House, is one of the major efforts aimed at addressing this challenge. In order to better understand the behavior of these materials and further design materials with targeted properties, researchers have relied on the materials science paradigm of processing-structure-property (PSP) linkage. Data-driven approaches founded on PSP linkage have much attention recently and have become a hot topic in many areas of materials research. Data-driven approaches combine materials science, computer science and statistics with the goal of expediating material discovery by utilizing past research data to draw new fundamental understanding. To facilitate the development of data archiving, sharing and development of data-driven approaches, efforts have been made to create online databases for fast queries and reference. A data resource, NanoMine, has been developed to accelerate material design for polymer nanocomposites. NanoMine allows fast data queries, visualization, and sharing, as well as a

number of tools for analysis including microstructure descriptor identification and reconstruction tools. There are three critical requirements needed to expand the functionality of NanoMine for usage by the broader nanocomoposite community. First, more analysis and module tools that better model the behavior of the material must be built. Second, case studies highlighting the capabilities of NanoMine must be developed using curated data to quantify PSP relationships and elucidate material mechanisms and physics. Finally, it is essential to put continuous efforts toward robust data curation, which expands the size of database and enables development systematic studies.

In relation to the three needs above, this dissertation first presents a combined finite element analysis (FEA) and optimization approach to accelerate the identification of interphase properties given experimental data of bulk nanocomposite properties. This approach is tested on both simulations of dielectric and viscoelastic properties in nanocomposites. Our work provides insight into identifying interphase properties for polymer nanocomposites using adaptive optimization and demonstrates the potential of data-driven approaches for achieving a deeper understanding of interphase properties that have proven difficult to directly characterize experimentally. Secondly, we present a novel deep learning approach that probes the structure-property relationships in polymer nanocomposites. Analysis of archived experimental data motivates the development of a computational model that allows for demonstration of this approach and gives flexibility to sufficiently explore a wide range of structures. Lastly, to facilitate the data curation process from literature sources, by applying recent machine learning and natural language processing methods, an end-to-end framework to extract material processing and synthesis information from full-length journal articles is developed. The proposed models and

methods are shown herein to be a powerful tool to supplement labor-intensive manual curation and improve the efficiency of data input for the database.

ACKNOWLEDGEMENT

It has been a great fortune spending the past five years to complete my Ph.D. study at Northwestern University. The first sincere appreciation goes to my Ph.D. advisor Dr. L. Catherine Brinson for her support, encouragement, guidance and criticism. I would also like to thank two other committee members Dr. Wei Chen and Dr. Linda Schadler for their suggestions to my research on multiple projects, some of which has been successfully published and listed as chapters here.

I would also to express my thanks to all the Brinson group members and alumnus at both Northwestern and Duke University. Thanks to Dr. He Zhao, who was the mentor when I joined Brinson Lab and walked me through basics of finite element analysis and brought me up to speed in independent research. Thanks to Dr. Xiaolin Li, Dr. Min Zhang (Theoretical and Applied Mechanics) and Mr. Zijiang Yang for all the discussions with insights for many ideas shown in this work. Thanks to all the other colleagues in Brinson group, Dr. Zhiwei Cui, Dr. Marc Palmeri, Dr. Pavan Kolluru, Dr. Ikumu Watanabe, Dr. Min Zhang, Dr. Partha Paul, Ms. Anqi Hu, Mr. Matt Eaton, Mr. David Collinson, Mr. Ridvan Kahraman, Ms. Anqi Lin and Mr. Bingyin Hu.

My appreciation also goes to my collaborators at Northwestern and Rensselear Polytechnic Institute. Thanks Dr. Yichi Zhang and Mr. Akshay Iyer at Northwestern for providing many helps on the microstructure characterization and reconstruction as well as design and optimization methods. Thank you to Dr. Aditya Prasad for providing experimental data on nanocomposite samples.

Lastly, I would like to express my sincere appreciation to my parents Mr. Xiling Wang and Ms. Xin Li for their generous support and belief in me. The most special thanks to my wife, Ms. Fanyu Zuo, who has always been supportive and by my side no matter how life goes up and down. Thanks to all the other family members and I could never achieve anything without their support and encouragement.

Table of contents

List of figures

List of Tables

Chapter 1. Motivation and research tasks

1.1. Background and motivation

Because of enhanced properties and the potential of being next generation materials, polymer nanocomposites have become a hot research topic in recent years [1-3]. Polymer nanocomposites have shown significant improvement in thermal, optical, mechanical, dielectrical and other physical properties compared with neat polymer systems, by adding exceptionally low filler to the neat polymer system [4-8]. Because of the existence of large extent of interphase - a polymer region with significantly modified dynamics – that results from both strong chemical and geometric interactions between the particle surface and the polymer segments near the particle (Figure 1, top right), in conjunction with the high surface-to-volume ratio of the nanoinclusions, the composites is able to show significant property enhancements even at low filler loading percentage. For example, a nanocomposite material containing 5 wt% of perfectly dispersed spherical nanoparticles with 40nm diameter contains a total interfacial surface area of about $3.5 m^2 / cm^3$ [9]. The significant extension of the interphase domain (Figure 1, top middle) is because of the interaction, cooperative nature of the macromolecular network, which leads the interphase area extends well beyond the matrix chains directly attached to the filler surface [10].

Because of their easy processability and outstanding properties, polymer nanocomposites have become more popular in industry. Figure 1 highlights some applications of polymer nanocomposite materials in daily life.

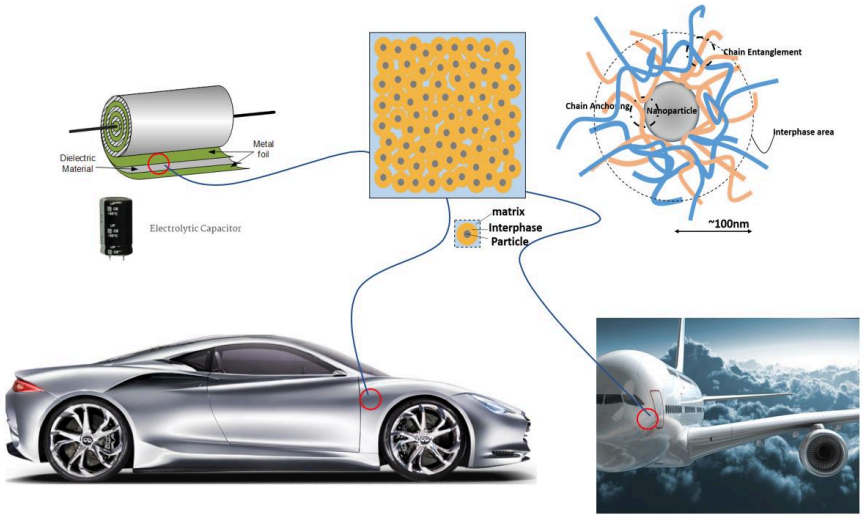

Figure 1. Industry application of polymer nanocomposites and the polymer nanocomposite interphase

Although advanced materials, including polymer nanocomposites, display tremendous advantages over traditional materials, the time frame from development of a new nanocomposite material to deployment of those materials into industrial applications is lengthy, typically over 10 years. Therefore, researchers and engineers seek to prioritize the reduction in time required to bring advanced materials from laborites to the market. The Material Genome Initiative (MGI), proposed by the White House, is one of the efforts aimed at addressing these challenges. The core of the MGI framework is to develop novel, integrated, material-innovation infrastructure leveraging computational, experimental, and data informatics methods to enable rapid engineering design, yielding a remarkable opportunity to replace traditional time-consuming design approaches mainly relying on trail and error [11]. Advances in computational modeling and simulations, including Finite Element Analysis (FEA) and Molecular Dynamics (MD) simulations, enable fast prediction of material behavior, as well as design validation. With advances in computational power,

researchers increasingly emphasize computational investigations. However, validation of key results using experimental data is essential to ensure that computational models are robust.

In order to better understand the behavior of nanocomposite materials and design new materials with tailored properties, researchers have relied on the materials science paradigm exploring processing-structure-property (PSP) relationships. Data-driven approaches founded on PSP linkage have attracted more attention recently and become a hot topic in many areas of material research in order to accelerate materials design. Data-driven approaches combine materials science, computer science, and statistics with the goal of enhancing materials discovery by drawing new connections using past research data [12, 13]. To facilitate data archiving, sharing and development of data-driven approaches, efforts have been made to create online databases for fast queries and reference. Representative works include CALPHAD database for alloy phase diagrams [14], Materials Project for inorganic compounds [15], and OQMD for complex alloys and compounds [16, 17]. For polymer nanocomposites, with the complexity of different types of data across various sources, it is hard to find a universe standard or schema that contains all the possible data coming from different perspectives. Several examples of online polymer data resources include the PolyInfo database from NIMS of Japan [18], the CRC POLYMERSnetBASE by Taylor and Francis Group [19], and the Polymer Property Predictor and Database (PPPDB) by University of Chicago. However, the complexity of polymer nanocomposites far surpasses even that of polymers due to factors such as the enormous variety of nanoparticle candidates and surface chemistries, the increase in processing complexity and options relative to pure polymer counterparts, and the interphase resulting from polymer-substrate interactions.

Despite the previous efforts toward developing deeper understanding of polymer nanocomposites, there are still some challenges remaining that hinder property prediction and material design, which will be discussed in Section 1.2.

1.2. Challenges and research tasks

1.2.1 Challenges in characterization of interphase properties

The interphase plays an essential role in the determination of nanocomposite properties due to the large surface area between polymer chains and the embedded nanoparticles. In order to understand the behavior of the interphase and make accurate predictions of nanocomposite properties, both experimental and computational efforts have been made [20-23]. Although direct measurements of the interphase in nanocomposites are challenging due to experimental limitations, some recent studies have focused on measuring the local, nanoscale elastic and viscoelastic properties of polymer regions near substrate surfaces using 'model nanocomposite' systems consisting of polymer films deposited on various substrates, shedding insight into local polymer property variations of confined polymers [24-29]. In particular, an atomic force microscopy (AFM)-based method measuring the local mechanical properties of polymers near to a substrate with nanometer resolution shows a gradient of mechanical properties extending tens to hundreds of nanometers from confining surfaces [30, 31].

To predict properties of nanocomposites, continuum mechanics methods can be employed with explicit interphase properties in the polymer regions surrounding nanoparticles [32-35]. In those continuum models, under some circumstances, the interphase properties can be described by

shifting factors based on the pure matrix properties, reflecting the modified dynamics of interphase regions. Given experimental data for bulk nanocomposite properties, the interphase properties can be determined from a trial-and-error based iterative tuning procedure by matching simulated results from FEA with experimental data, using some basic assumptions about the length scale, magnitude, and profile of the modified interphase properties [36-39]. However, there are several disadvantages for a trial-and-error based manual fitting method. First, this process can be tedious and time-consuming given the computational cost of FEA and the complexity of experimental data. For example, a single FEA model for one sample with ~10k elements requires 30 minutes to run on a typical server. Given the fact that the optimal shifting factors at each iteration are guessed based on the previous results and there are many iterations required to obtain optimal results, this manual tuning process provides tremendous challenges for efficient investigations of the correlations between the interphase properties and the material constituent characteristics. Automating the iterative process, however, offers a more efficient route toward determining the interphase properties needed to match simulations with observed experimental results, thereby helping draw deeper understanding of the PSP relationships.

1.2.2 Challenges in structure-property relationship prediction

To predict the properties of nanocomposites with given dispersion states, or morphologies, the traditional approach depends on numerical simulations such as continuum mechanics methods and multi-scale simulations. A variety of micromechanical models such as Mori-Tanaka, Halpin-Tsai and the self-consistent scheme have been developed to predict the thermomechanical behavior of nanocomposites [32-35]. These analytical models are insufficient to fully capture the true

morphology of the fillers in real nanocomposites, although some models include structural parameters. Therefore, multi-scale simulations are often necessary. Finite element (FE) simulations are able to more accurately capture the structural information and accommodate non-homogenous material systems with explicit configurations for all material phases, making FE simulation a strong candidate to analyze the behavior of nanocomposites.

Although simulation methods are often relatively quick and inexpensive compared with experimental testing and are often less hindered by some practical experimental limitations, the computational cost still can be expensive, especially for atomic-scale simulations such as molecular dynamics, which are often limited to small regions, which hinders the exploration of the bulk properties of multiscale materials (e.g., nanocomposites with non-uniform dispersion states). Therefore, it is necessary to build metamodels or surrogate models representing the input and output relations in simulations. In order to build such surrogate models for predicting structure-property relationships, a sufficiently representative pool of data is required. Traditional approaches building such linkage usually depend on designing hand-crafted features from microstructures by domain experts that sufficiently capture true morphology characteristics. Because of using hand-crafted features, the accuracy of this approach depends heavily on the quality and resolution of designed features and selecting the most important features from a large pool of candidates is nontrivial. Moreover, the designed features are usually nontransferable from one material system to another. For example, the features selected to characterize polymer nanocomposites may not be useful to represent many classes of metamaterials. Therefore, to explore structure-property relationships in polymer nanocomposites, targeted data generation and an end-to-end, interpretable machine learning model with minimal human intervention is required.

1.2.3 Challenges in data acquisition from published literature

Although data-driven material platforms such as NanoMine provide a path for rapid material design and development, data curation remains a challenge, as it mainly relies on manual data extraction from scientific literature sources by experts with domain knowledge. The manual curation procedure is considered to be expensive, labor-intensive and error-prone. For example, based on our experience during the infancy of NanoMine, a curator often requires several hours to extract all of the processing-structure-property information from one paper and input the data into the database. Moreover, with an enormous number of previously published studies on polymer nanocomposites and continuing publication of new results, it is becoming increasingly difficult to maintain and update manually curated databases.

There are many well-established text-mining tools implemented in the biomedical and chemical domains [40-45]. Most of these tools can extract specific entity types from specific document domains, whereas some of them can be utilized for a broader focus on extraction of chemical information, experimental processing, properties and even relationships between entities. One such tool is called ChemDataExtractor [44], which is a comprehensive toolkit for the automated extraction of chemical information from the scientific literature. This toolkit offers a complete natural language processing (NLP) pipeline for the automated extraction of chemical information from published sources. Making use of a wide range of state-of-the-art methods, including part-of-speech (POS) taggers, word clustering, and name entity recognizers that combine conditional random field and dictionaries, this toolkit is able to automate the extraction of chemical entities, properties, characterization techniques, and processing procedures. Additionally, ChemicalTagger [45] has been developed to parse experimental processing sections of the

documents to determine the chemical roles (e.g., reactant, solvent) and experimental processing procedures (e.g., heating, mixing methods, etc.) by using rule-based text parsing.

However, the field of polymer nanocomposites is lacking a tailored NLP tool that is able to extract relevant material information from the literature. Additionally, given the fact that most of the NLP tools are domain specific, the tools developed in other fields usually have poor performance if directly applied to nanocomposite publications. Therefore, it is critical to build NLP tools that can automatically extract material information from the literature, thereby vastly enhancing the speed and reliability of the data curation process to expand the large, robust databases needed for data-driven material platforms.

1.3. Research tasks

NanoMine serves as a data-driven platform for polymer nanocomposites that includes a database, search and visualization tools, microstructure reconstruction tools, and physics-based simulation tools [46, 47]. As shown in Figure 2, there are three major g of NanoMine: the first is to develop data-analysis and simulation tools that better capture the physics of nanocomposite behavior and enable deeper understanding of the mechanisms behind observed material behaviors; the second is to build data-driven machine learning and design models to draw new insights into PSP relationships using data stored in NanoMine; the last is to curate sufficient data containing material processing, structure and property information for the aforementioned analysis. The research tasks described in this thesis correspond to each of these three objectives, as described below. The first task focuses on improving the accuracy and robustness of computational models by developing an automated ‘inverse method’ to identify the interphase properties in polymer

nanocomposite using adaptive optimization. The method allows for computational characterization of interphase properties that can be validated through comparison to corresponding experimental results, enabling correlation between interphase properties and material constituents subjected to various processing conditions to fabricate nanocomposites. The second task utilizes data and tools in Nanomine to build data-driven models to identify structure-property relationships for polymer nanocomposites. The last task is to implement automated data curation protocols using NLP techniques to robustly extract data from publications into NanoMine. Details about each task are elaborated as follows:

Task 1: An inverse approach to identifying interphase properties in polymer nanocomposites: In this investigation, a combined FEA and optimization approach to accelerate the characterization of interphase properties given experimental data of bulk nanocomposite properties is presented. The objective of this task is to find the optimal interphase properties that minimize the difference between simulations and corresponding experimental results using an automated procedure.

Task 2: Mining structure-property relationships using NanoMine: Structure-property relationships in polymer nanocomposites are investigated using both qualitative analysis and quantitative methods. This study presents an analysis of experimental data from NanoMine, which motivates extension to complementary analysis of computation data, giving flexibility to explore a wider range of structures that are challenging to investigate experimentally. The objective is to demonstrate the power of NanoMine in finding the underlying physical mechanisms of nanocomposite behavior and to guide the design of new nanocomposite materials with tailored properties.

Task 3: Extracting material synthesis information from the scientific literature using NLP: In order to accelerate robust data extraction, reduce possible errors, and provide more insight into nanocomposite material design, an advancement toward fully automated data curation using machine learning is presented. By applying recent statistical learning and natural language processing methods, this task is aimed to develop an end-to-end framework to extract synthesis and processing information for polymer nanocomposites from full-length journal articles.

1.4. Dissertation outline

The outline of this dissertation is given in Figure 2. Chapter 1 contains an introduction to polymer nanocomposites and challenges in the design of new nanocomposites with tailored properties, which motivates the research tasks in this dissertation. Chapter 2 provides technical background for the investigations presented in subsequent chapters. Corresponding to the three critical requirements for NanoMine expansion and broad implementation by the polymer nanocomposite research and design community, three research tasks are proposed in Chapters 3-5. Chapter 3 presents the development of an inverse method to identify interphase properties from experimental data of bulk nanocomposite properties using adaptive optimization. Chapter 4 describes a case study of the capabilities of the NanoMine data resource: mining structure-property relationships in polymer nanocomposites. Chapter 5 introduces a semi-automated framework applying natural language processing (NLP) to enhance the curation of polymer nanocomposite data from scientific literature. Chapters 3-5 additionally include perspectives into further extending the functionality of NanoMine for advanced material design. Chapter 6 summaries the

contributions of the work presented in this dissertation and points to potential future directions to build upon the findings.

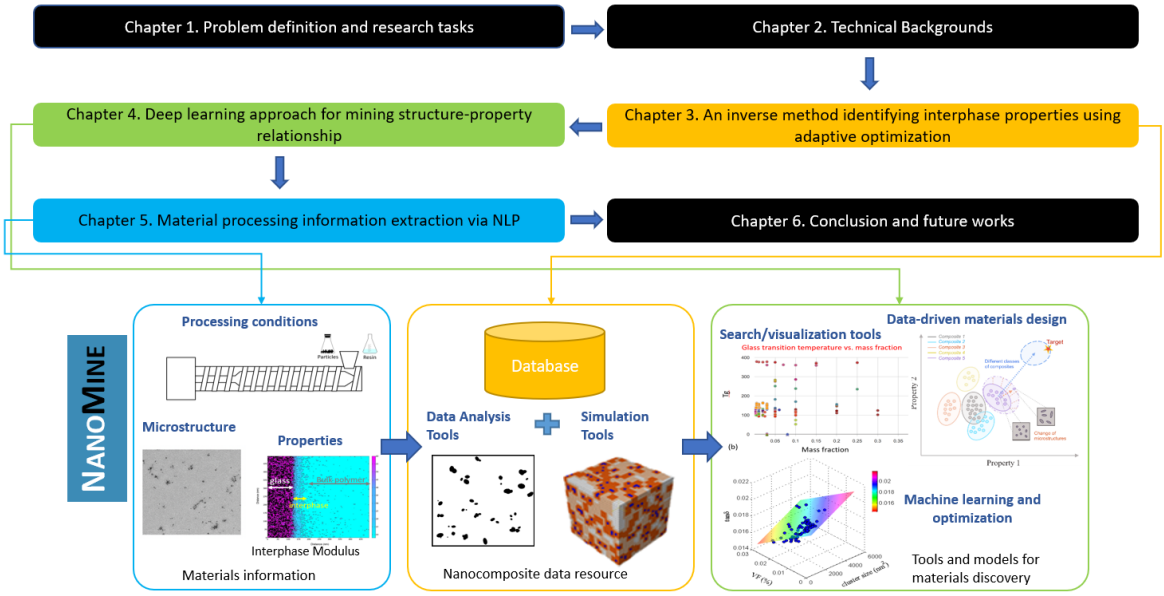

Figure 2. NanoMine consists database, data analysis tools and simulation tools for data-driven material deign (bottom). The outline of this paper (top).

Chapter 2. Technical background

This chapter provides technical background underlying the concepts discussed in this dissertation. In section 2.1, a review of structure-property modeling approaches, including both numerical approaches and data-driven approaches, are reviewed. Section 2.2 presents the current status of techniques used to determine the interphase properties in polymer nanocomposites, including both direct measurements from experiments and simulation methods. Section 2.3 provides a basic introduction to deep learning, which is applied to build structure-property linkage for task 2 and NLP models for task 3 in Chapters 4 and 5, respectively. Some popular basic

architectures including convolutional neural networks for image processing tasks and recurrent neural networks for language modeling tasks are introduced.

2.1. Structure-property modeling

To accurately determine structure-property relationships in materials, numerical models, including analytical methods, continuum level simulations, and atomistic level simulations have been developed. In recent years, data-driven approaches combining machine learning models with materials data have also become popular. Each of these methods are presented in this section.

2.1.1 Numerical approach

Numerical approaches typically model polymer nanocomposites as multi-phase materials with three distinct phases: polymer, particle and interphase [9, 27, 32, 33, 48, 49]. Both analytical methods and simulation models such as FEA simulations and atomistic level models are developed to predict various properties of composite materials, including thermomechanical and dielectric behavior.

Analytical methods: micromechanical models, such as Halpin-Tsai, Mori-Tanaka and the self-consistent scheme have been applied to predict the thermomechanical properties of composites [50]. A variety of analytical models have been developed to analyze the dielectric behaviors including the Bruggeman model [51], Lichtenecker model [52] and Todd-Shi model [36]. While numerical approaches have great advantages in terms of time efficiency relative to simulation-based methods, most analytical models do not fully account for structural information, reducing prediction accuracy.

FEA simulations: FEA simulations can accurately represent complex, non-homogeneous material systems with explicit configuration of all relevant material phases. This makes FEA a good option for analyzing the behavior of nanocomposites, incorporating both important nanofiller dispersion and interphase properties. Finite element models for analyzing the thermomechanical and dielectric behavior of polymer nanocomposites have been utilized to investigate the impact of interphase on bulk nanocomposite properties [27, 29, 33, 37, 49, 53-56].

Atomistic level simulations: Atomic-scale simulations, such as molecular dynamics (MD) and density function theory (DFT) computations, have been implemented to model the mechanical and dielectrical behavior of nanocomposites [57, 58]. In polymer nanocomposite field, MD/DFT provides fundamental understanding at the nano-scale of intrinsic interphase properties and to bridge scales from atomistic level to continuum level [59, 60].

2.1.2 Data-driven approach

Although numerical simulations provide a method to predict structure-property relationships without the need to perform tedious experimental investigations, the computational cost of simulations can be expensive, especially for large volumes of complex, heterogenous material systems. This poses a significant challenge for exploring the vast design space for materials such as polymer nanocomposites. Data-driven methods are emerging to solve such problems by constructing 'models of the models' via data mining and machine learning techniques. In recent years, data-driven methods have gained increasing interest for predicting structure-property relationships in various types of materials, including nanocomposites [17, 60-62]. Most data-driven approaches rely on design of hand-crafted features that capture the characteristic of

microstructures, followed by implementation of machine learning models to correlate the extracted features with target property values. Approaches using geometric descriptors and correlation functions are detailed below:

Geometric descriptor approach: In order to illuminate structure-property relations in multi-phase materials, one of the traditional methods entails building correlations between user-selected features found to have the most impact on material behavior. Those features are usually chosen to serve as a low-dimensional geometric representation of the material and can include parameters such as composition, morphology, dimensions of discrete particles, etc. In this process, the minimum number of significant features that accurately reflect the modeled material are selected for subsequent machine learning of structure-property relationships, with accuracy of the technique largely determined by how well the chosen parameters capture the microstructure of the modeled material. The selected features generally depend on the specific material system. For polymer nanocomposites, geometric descriptors such as volume fraction, aspect ratio of nanoparticles, and nearest neighbor distance between nanoparticles are often critical parameters for controlling the material response [63-65]. On the other hand, for metallic materials, features such as density and size of voids, grain size, and average length of grain boundaries are often used [66]. A regression model such as linear regression or random forest regression is applied to find the influence of each geometric feature on resultant properties. This method has high interpretability since the impact of physical features of the system on material behavior can be easily calculated. However, this strategy is limited by a lack of generality and accuracy for different classes of microstructures.

Correlation function approach: Polymer nanocomposite microstructure can also be characterized using statistical functions. The N-point correlation function can be used to quantify the dispersion state and contains richer information than the geometric descriptor approach, as no averaging is implemented [67-70]. In practice, the two-point correlation function is typically calculated, followed by subsequent dimension reduction using principal component analysis (PCA) to reduce computational cost, as two-point features lie in a high-dimensional space [66, 71]. For example, a microstructure with a size of 256*256 pixels has a two-point correlation of dimension 256*256. Due to the dimension reduction afforded by PCA, the correlation function approach can also be considered to be a method using hand-selected features. The top n principle components are applied to the regression model to predict the target property.

2.2. Nanocomposite interphase modeling

The interphase in polymer nanocomposites is an explicit volume of polymer near an interface with modified properties relative to those of the bulk polymer and consists of two regions: the extrinsic interphase (order ~1-10nm), whose properties is determined by the extrinsically placed molecules on the filler surface and chemical interactions at the interface, and the intrinsic interphase (order ~100nm), whose properties come from changes in mobility or structure of the polymer from interactions with the nanofiller surface [72]. The interphase often plays a critical role in determination of bulk properties of nanocomposites due to percolation of interphase regions through the microstructure. Efforts have been made both experimentally and theatrically to understand the underlying mechanisms underlying interphase behavior in order to better predict nanocomposite properties.

Experimental Approach: Although direct measurement of the interphase is challenging because of limitations in experimental techniques at the nanoscale, some recent studies have focused on measuring the local elastic and viscoelastic properties of polymer regions near interfaces using idealized 'model nanocomposite' systems of polymer films on flat substrates with well-defined geometries that can shed insight into real nanocomposite systems with more complex morphologies, indicating that the local polymer properties are significantly altered in the vicinity of polymer surface [24-26, 28-30]. In particular, an atomic force microscopy (AFM)-based indentation method directly probing the local mechanical properties with nanometer resolution shows a gradient of mechanical properties extending approximately tens or hundreds of nanometers from confining surfaces before reaching bulk values farther from interfaces [30, 31].

Furthermore, electrostatic force microcopy (EFM) can be used to measure the local dielectric permittivity of the polymer interphase near surfaces. In this testing mode, local interphase areas are probed using a conductive tip with an AC voltage. The tip is kept at a relatively large distance from the sample in order to eliminate the effect of short-range atomic forces and only the long-range electrostatic force is measured. The local permittivity of the material can be derived by analyzing the shift in resonant frequency [73].

Numerical Approach: Different models have been developed to directly simulate interphase properties or to help interpret empirical data from experiments. Coarse-grained molecular dynamics (CGMD) is applied to simulate the AFM indentations and to investigate the elastic modulus near interfaces [59]. On the other hand, although direct measurement of interphase properties is achievable via AFM or EFM, the data from mechanical indentations are usually affected by the interaction of the elastic stress field from the tip interacting with the substrate

during indentation. FEA modeling complements these experiments by simulating the indentations in order to help deconvolute the influence of the substrate stiffness, tip radius, and other factors on the measured data so that interphase properties can be extracted [74]. A 3D FEA indentation model can be used to interpret experimental indentation data in order to recover the true gradient profile of the interphase modulus.

2.3. Deep learning

In recent decades, machine learning has played an increasing role in many aspects of modern society: from web-search engines to news-feed ranking on social networks to recommendations on e-commerce platforms. Machine learning has even impacted other industries such as medical science, chemistry and materials science [75, 76].

Conventional machine learning usually requires feature engineering and domain expertise to extract critical features of a system into vector form in order to represent raw data for subsequent analysis. This greatly restricts the ability of conventional machine learning strategies to process the data in their raw form. To solve this problem, deep learning methods have been proposed with multiple levels of internal representation using non-linear functions that directly transform the raw data into a higher and more abstract level. Deep learning has greatly improved the previous benchmarks in image recognition and speech recognition. Moreover, deep learning has beaten other machine learning models in materials science applications, such as microstructure reconstruction, property prediction, and material design.

A basic architecture of multi-layer neural networks used in deep learning is shown in Figure 3**.** A deep learning model may consist of many hidden layers that are able to capture

complex non-linear relationships between input and output. The function of the hidden layer is to transform the input in a non-linear manner such that the output layer works as a linear regression or classification model. To calculate the values for each layer, we first compute the value z – the total weighted sum of the values from the previous layer. Then, a non-linear activation function $f(\cdot)$ is applied to z and generate outputs for each layer. Typical activation functions include the rectified linear unit (ReLU), tanh, and sigmoid. A deep learning model may consist of millions of trainable parameters, which are updated using a backpropagation algorithm. The backpropagation procedure is an application of a chain rule to compute the gradient of the objective function with respect to weights at each layer.

There are different types of deep neural networks: feedforward networks, convolutional neural networks (ConvNet), recurrent neural networks, and generative adversarial networks. The architecture of feedforward networks is similar to Figure 3**.** In this thesis, convolutional networks and recurrent neural networks are applied. Further details of these two architectures are described in the following subsection.

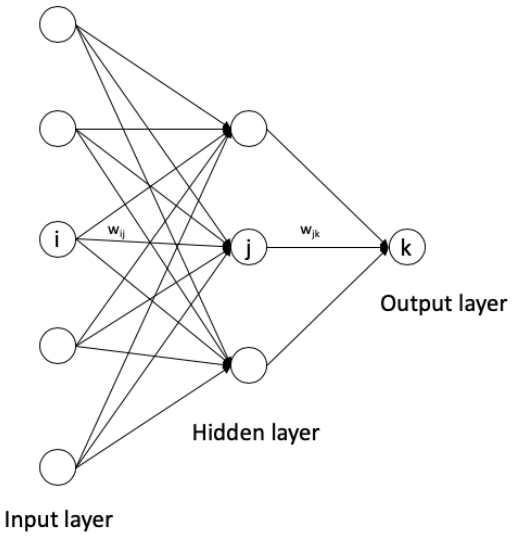

Figure 3. Basic architecture of an artificial neural network

2.3.1 Convolutional neural networks

Convolutional neural networks (ConVNets) are built to process data in the form of arrays, such as signals, audio data, and images [77, 78]. The architecture of a basic ConVNet is shown schematically in Figure 4.

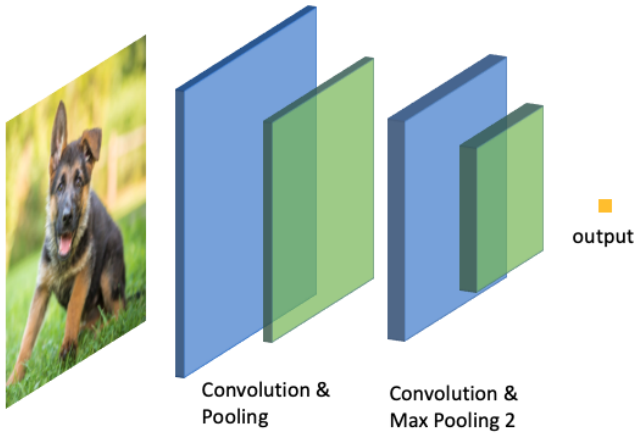

Figure 4. Basic architecture of a convolutional neural network

The basic building units of ConVNets are convolution and pooling layers. The convolution operation is done by calculating the weighted sum between the convolution filter (also called filter) and the portion of the image which overlap with the position of the filter. A filter much smaller than the image (3*3*3 pixels, for example) slides across the entire image using a specified step size during processing, as shown in Figure 5**.** The role of the convolution layer is to detect local correlations from the previous layer.

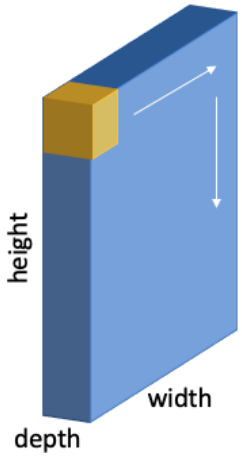

Figure 5. The movement of the filter through the image.

The pooling process is designed to further reduce the dimensionality of feature maps and merge semantically similar features into one. Common pooling processes include max pooling or average pooling, which are able to reduce the dimensionality by computing the maximum or average value, respectively, of a local patch of units in one feature map. By stacking convolution and pooling layers, a deep ConVNet can be built that is able to extract higher order abstract and complicated features in a stepwise manner.

ConVNets have been widely applied to process array-like data such as signals, language, and images. Among all of their applications, ConVNets have great success in image processing

tasks such as the detection, segmentation and recognition of objects and regions. In recent years, the ConVNets entail deeper processing, with most architectures having 10 to 20 layers with millions of trainable weights and billions of connections between different units [79, 80]. Because of the advances in hardware, software, and algorithms, the training time has been dramatically reduced from several weeks to several hours.

2.3.2 Recurrent neural networks

Recurrent neural networks (RNNs) work better for data with sequential inputs such as speech and language. RNNs have successful application in machine translation, language generation, and image captioning. Unlike ConvNets, RNNs process the input data one element at a time, keeping a 'state vector' of that specific element in the hidden units such that the model explicitly captures the sequential information about the data [81-83]. A typical RNN architecture is shown in Figure 6.

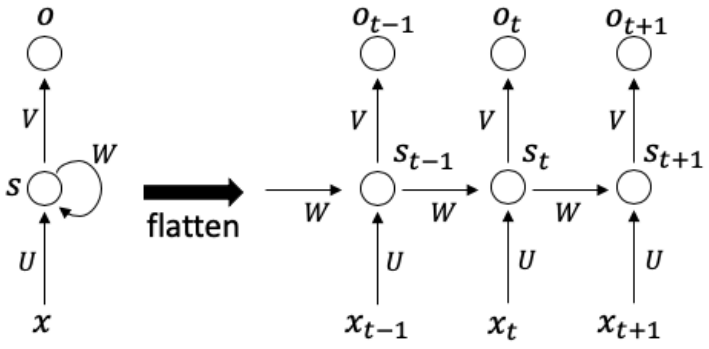

Figure 6. Basic architecture of a Recurrent neural networks (RNNs)

In a RNN, each neuron not only takes input from the current sequence x_t but also from other neurons at previous steps. In this way, the output o_t depends on the inputs from all previous steps, and the sequential information from the input is captured by the model. The RNNs can be seen as a very deep feedforward neural network where all the layers share the same set of weights $(\mathrm{U}, \mathrm{V}, \mathrm{W})$. One of the difficulties in training RNNs is the gradient vanishes or explodes, especially when the sequence is long. This is because the backpropagated gradient either grows or shrinks at each step, and if there are many steps (i.e. the sequence is long), the gradient will vanish or explode. To solve this problem, long short-term memory (LSTM) networks with special memory cells are designed [83]. The special memory cell is able to learn and determine when to clear the long-time dependencies. LSTM has been shown to be more useful than conventional RNN especially for the case of long sequences.

Chapter 3. Identifying interphase properties in polymer nanocomposites using adaptive optimization[1]

3.1. Introduction

As mentioned in chapters 1 & 2, Polymer nanocomposites have attracted great interest in recent years because of their potential as tailored materials with enhanced properties [1-3]. One of the reasons for the enhancement in properties is the large interphase region that results from both the strong chemical and geometric interactions between the particle surface and the polymer segments near the particle and the high surface-to-volume ratio of the nanoinclusions. For example, if a composite sample contains 5 wt% of 40nm particles perfectly dispersed, the resulting total interfacial surface area is about $3.5 m^2 / cm^3$ [9]. As illustrated in Figure 7, due to the interacting cooperative nature of the macromolecular network, the interphase area extends beyond the layer of matrix chains directly bound to the filler surface resulting in the significant extension of the interphase domain into the matrix [10]. In order to understand the behavior of the interphase and make accurate predictions of nanocomposite properties, efforts have been made to measure the interphase thickness and its mechanical or dielectric response [20-23]. Although direct measurements of the interphase are limited because of challenges in experimental visualization at the nanoscale, recent studies focus on measuring the local elastic and viscoelastic properties in different polymer microdomains by correlating thin film and nanocomposite data, providing

[1] This chapter is from published paper on *Composites Science and Technology* 162 (2018): 146-155

adequate evidence that the local polymer properties are significantly altered in the vicinity of polymer surface [24-29]. In particular, an atomic force microscopy (AFM) -based method that directly measures the mechanical properties of polymers adjacent to a substrate with nanometer resolution shows a gradient of mechanical properties extending approximately 100nm from confining surfaces [27, 84].

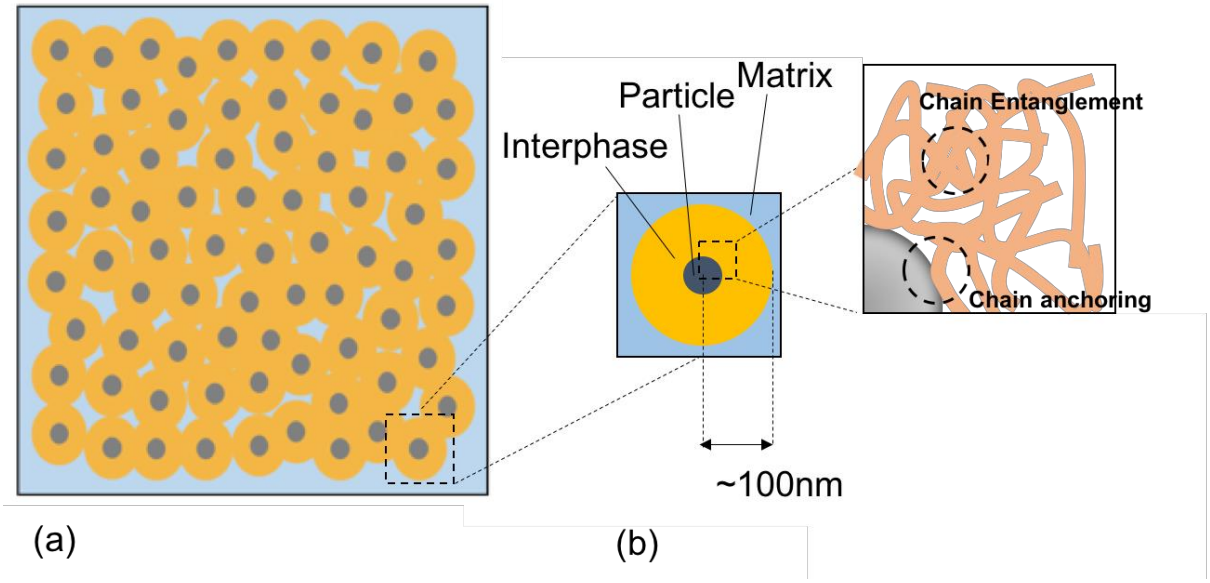

Figure 7. Interphase configuration. (a) Schematic of the interphase regime in a nanocomposite sample. (b) Schematic showing the extended interphase structure. Yellow chains directly anchored to or chemically interacting with nanofiller. Blue chains impacted by filler indirectly through interaction with yellow chains. Cooperatively interacting chains propagate interphase zone to persist to the order of 100nm from the particle surface. Note: figures not to scale

To predict properties of nanocomposites, continuum mechanics methods are often employed in which three phases must be considered: polymer, particle and interphase. Micromechanical models [50], such as Halpin-Tsai, Mori-Tanaka and the self-consistent scheme have been applied to predict the thermomechanical behavior of nanocomposites [32-35]. A variety of analytical models have been developed to analyze the dielectric behaviors including the Bruggeman model [51], Lichtenecker model [52] and Todd-Shi model[36]. Given experimental

limitations to measure interphase properties directly, one approach to determine the interphase properties is inversely through tuning the parameters in finite element analysis or micro-scale model constitutive equations using the bulk composite properties from experiments [36-39]. Importantly, in order to capture the dispersion state or the morphology information of the fillers, multiscale simulations are often necessary. For example, multiscale approaches have been applied to study the viscoelastic properties in polymer nanocomposites [27, 60, 85].

FEA simulations can accommodate complicated non-homogeneous material systems with explicit configuration of all relevant material phases. This makes FEA a good option for to analyze behavior of nanocomposites and include both important nanofiller dispersion as well as interphase properties. We have developed finite element models for analyzing the thermal and mechanical [37, 48, 53, 86] and dielectric behavior [38, 87] of polymer nanocomposites and investigated the impact of the interphase on the corresponding properties. We have shown that in some cases, the interphase properties can be described by shifting factors based on the pure matrix properties, which can be well represented by the Prony Series as a parametric expression of multiple relaxation times and strengths. Given experimental data for both the pure matrix and bulk nanocomposites properties (either dielectric or thermomechanical) the necessary interphase properties can be determined from a trial-and-error based iterative tuning procedure by matching simulated results from FEA with experimental data. However, there are several disadvantages for this trial-and-error based manual fitting. First, this process can be very time-consuming given the complexity of experimental data and computational cost of the FEA (a single FEA model for one sample with ~10k elements requires 30 minutes to run on typical server). Since many manual iterations are often required, where the optimal shifting factors are guessed based on the previous outputs, this

tedious manual tuning process prohibits efficient investigation of the correlation between shifting factors and material constituent characteristics.

In this work, we present a combined FEA and optimization approach to accelerate the search of optimal interphase properties given experimental data of bulk property for the composite. Our objective is to find the optimal interphase properties that minimize the difference between simulations and experiments, and to do so with an automated procedure. We adopt an adaptive global optimization approach that incorporates Gaussian Processing (GP) modeling [88] and sequential sampling strategy [89] to efficiently find the global optimal solution. Our proposed method can accurately find the optimal shifting factors given experimental data in tens of iterations, which significantly eases the computation costs from simulations. We demonstrate our method by finding both dielectric and mechanical properties of the interphase based on composite property data. This method is an efficient and reliable tool to determine interphase properties and can facilitate future work of uncovering the relationship between interphase properties and material constituents.

3.2. Methodology

Our goal is an optimization of interphase properties for a single sample, for which we have a) constituent properties, b) composite properties, c) microstructure information. We seek to determine the interphase properties that will yield the composite properties (b) from (a) and (c). We assume the interphase properties can be represented by shift factors with respect to matrix properties (defined in more detail below). Thus, we seek to find the shift factors that will optimally match the composite data via an automated procedure applying adaptive optimization. Based on

experience and literature data we begin with reasonable bounds for the shift factor values, and the space defined by the n-dimensional factors (n=5 for dielectric case [87], n=2 for mechanical case[60]) varying in these bounds defines the search space for the adaptive optimizer. The key components of the adaptive optimization method are summarized in Figure 8: (1) The empirical bounds of shift factors are used to set the range to sample initial training sets of shift factors from design of experiments (DOE) using Optimal Latin hypercube (OLHC). For each set of shifting factors, the FEA model is run and outputs the simulated result. Then the objective function is formulated as the difference between the experimental data and the simulation using mean square error (MSE); (2) A surrogate model, in our case, a Gaussian Process (GP) model, uses the training data to learn the relationship between the objective (Difference F) and features (shifting factors), with uncertainties.; (3) adaptive optimization (selector) provides the most promising candidate points for the new simulation and augments the initial shift factor set from DOE. In our study, new candidate points are selected based on the feedback from the surrogate model at a previous step by calculating the Expected Improvement (EI). Step (2) and (3) is usually identified as an adaptive optimizer, which augments the initial training shift factor set and drives the subsequent iterative improvement of the surrogate model and prediction accuracy.

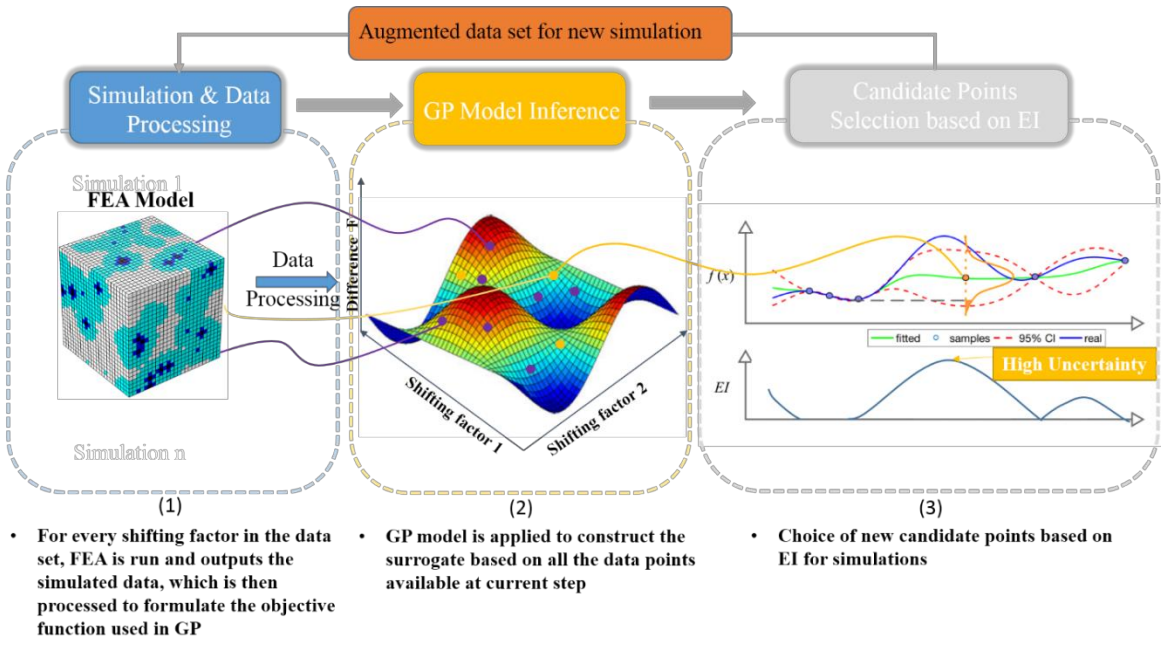

Figure 8 Our automated adaptive optimization strategy for searching the interphase properties: (1) FEA is run on each point in the initial shift factor set from DOE and then processed to formulated our objective function by calculating the difference between simulated results and experiments using MSE. (2) A GP model is applied to construct the surrogates for predicting the value of objective function (Difference F) within the design space. The purple points are initial sampling points from DOE and the yellow points are sequentially generated by EI in (3). (3) EI chooses the best candidate points for additional simulation. The new candidate points augment the initial training shift factor set from DOE to further improve the surrogate model and prediction accuracy.

3.2.1 Interphase FE model

Our method is developed and tested on two cases for polymer nanocomposite response data: viscoelastic simulations and dielectric simulations. For both cases, we have developed finite element models to simulate the composite properties given the microstructure from Scanning Electron Microscope (SEM) or Transmission electron microscopy (TEM) images [38, 48]. The schematic of the FEA model for dielectric studies and viscoelastic simulations is similar except different software is used to run the simulations (*COMSOL* for dielectric and *ABAQUS* for viscoelastic simulations). The FEA configuration is shown in Figure 9. The microstructure

dispersion information is directly obtained from TEM micrograph images Figure 9a). Then a binary image showing filler and matrix phases is generated by application of a previously developed Niblack analysis algorithm [90]. The binary image is then characterized to identify the geometric descriptors and reconstruct the equivalent 3D microstructure and assign each point to a material constituent in the FEA model (filler, interphase, matrix) (Figure 9c). The geometric information of the filler and matrix can be directly obtained from the original image, while the interphase regime between the filler and particle is assumed to be represented by an extended layer around each particle. Thickness of the interphase is reasonably assumed to be 50nm in our study based on previous studies on filler-filler spacing and experiment measurements [9, 27]. The FEA model is then run in the respective software to obtain the frequency-dependent viscoelastic or dielectric response.

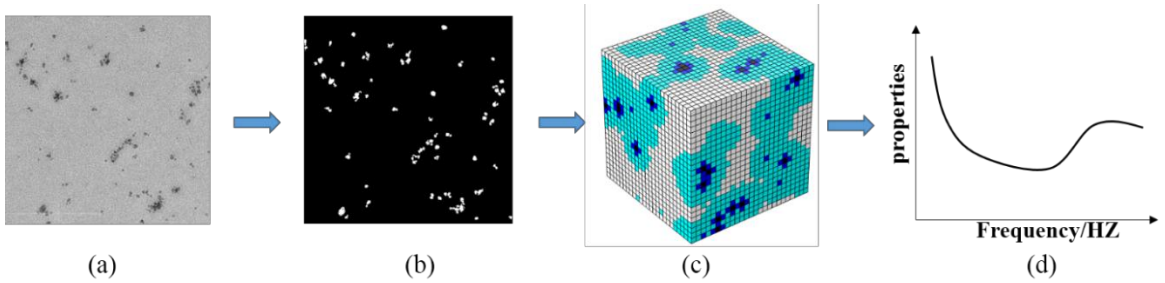

Figure 9 Configuration of FEA model in COMSOL (dielectric simulation) and ABAQUS (viscoelastic simulation). a) Sample TEM image, b) binary image of matrix and filler phases, c) reconstructed microstructure, d) frequency dependent dielectric properties or viscoelastic response predicted from simulation.

These interphase FEA models have been used with a manual inverse approach to determine interphase properties [48, 60, 87]. In these studies, the shift factors to relate interphase properties to matrix properties are defined differently for each property case. In viscoelastic simulations, a broadening factor B and frequency-shifting factor S are applied to describe the interphase behavior based on the matrix property. B accounts for the broadening effect of the loss peak while

S accounts for the horizontal shift of relaxation times in the frequency domain (Fig 4a). To describe the dielectric interphase, due to the multiple relaxations observed in the pure matrix, the alpha and beta relaxations are modeled separately and a five-dimensional shifting factor set is applied in the interphase model. The threshold value of the relaxation time is determined based on experimental data ($\tau_0 = 0.01$ in our case) so that the corresponding threshold frequency separates the relaxation peak into low (α relxation) and high frequency regions (β relxation). For each relaxation (alpha or beta relaxation), two shifting factors (M_a and S_a for alpha relaxation; M_β and S_β for beta relaxation) are applied to account for the change of intensity (M_a and M_β) and the shift of relaxation times (S_a and S_β) for respective relaxations. An additional constant, C, describes the intensity shift of ε'.

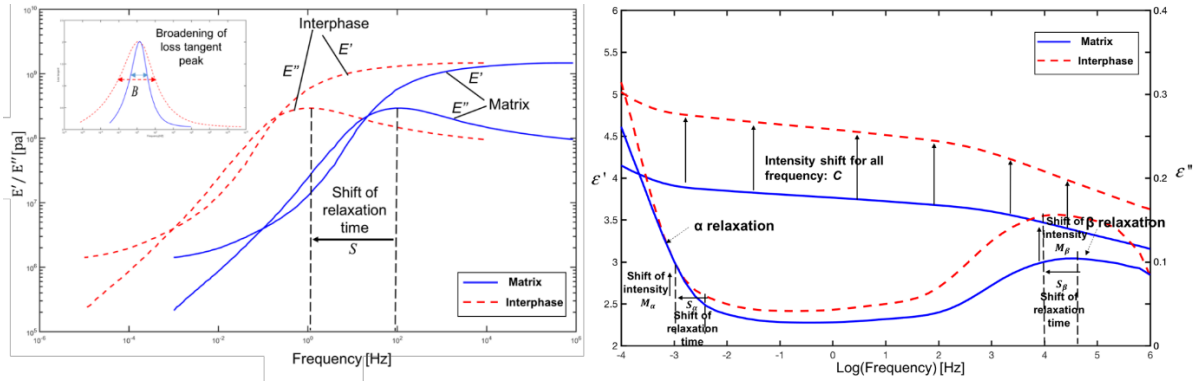

Figure 10. Relation of material properties between the interphase and pure matrix represented by shifitng factors in frequency space for (a) viscoelastic interphase properties with $S = 2, B = 2$ (b) dielectric interphase properties with $s_\alpha = 2, M_\alpha = 1.5, S_\beta = 2, M_\beta = 1.5, C = 0.5$

3.2.2 Optimization Objective

The objective of this work is the development of a rigorous automated method to predict the optimal interphase shifting factor set which enables a simulation result to best fit with the experimental data, thus determining the interphase properties is an inverse problem. There are a number of choices of mathematical measures that quantify the discrepancy between simulation and experimental results. These measures can be divided into three categories: magnitude-phase-comprehensive metrics, single-value metrics, and analysis of variance metrics. We tested the performance of different metrics on characterizing the difference between simulation and experimental results and found that the mean square error (MSE), which is one of single-value metrics, is the best descriptor for our problem. The MSE of the predictor in our study can be expressed as,

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(c_i - e_i)^2 \tag{1}$$

where c_i and e_i represent each simulated data point in frequency and experimental data point in frequency respectively.

In order to search for the optimal shifting factors, a proper target function needs to be defined. Ideally, the optimal shift factors should lead to well-matched data curves between FEA and experimental results. Equivalently, the minimal discrepancy between simulated and experimental data across the entire frequency band is desired. Additionally, for our case, it is required to minimize the difference from both real and imaginary property data simultaneously,

which leads to a multi-objective optimization problem. To addresses this problem, we first characterize the difference from the real and imaginary parts separately using Equation (1) for every shift factors set across the OLHC space (initially generated from DOE) and obtain the vector of real part difference $\mathbf{f}^1$ and imaginary part difference $\mathbf{f}^2$ for each data point in the OLHC space.

To develop the multi-objective approach, for every every shift factors set across the OLHC space we normalize the objective function as follows:

$$\mathbf{f}^{\mathbf{i}\_\mathbf{norm}} = \frac{\mathbf{f}^{\mathbf{i}} - \mathbf{f}^{\mathbf{i}}_{\min}}{\mathbf{f}^{\mathbf{i}}_{\mathbf{max}} - \mathbf{f}^{i}_{\min}} \tag{2}$$

Here, $\mathbf{f}^{\mathbf{i}}$ represents the vector of difference calculated from MSE for real (i=1) or imaginary part (i=2), $\mathbf{f}^{i}_{\min}$ and $\mathbf{f}^{\mathbf{i}}_{\mathbf{max}}$ represent the minimum and maximum difference value for the entire design space respectively. So from this equation, we are able to generate two normalized difference vectors for the real ($\mathbf{f}^{1\_\mathbf{norm}}$) and imaginary properties ($\mathbf{f}^{2\_\mathbf{norm}}$) separately. Then, we formulate this multi-objective function as the summation of weighted difference of the real and imaginary components:

$$\mathbf{F} = w_1 \cdot \mathbf{f}^{1\_\mathbf{norm}} + w_2 \cdot \mathbf{f}^{2\_\mathbf{norm}} \tag{3}$$

Here, w_1 and w_2 are the weight of real and imaginary components. $\mathbf{f}^{1\_\mathbf{norm}}$ and $\mathbf{f}^{2\_\mathbf{norm}}$ are normalized error vector from real and imaginary components, respectively. $\mathbf{F}$ is the final vector form of the objective describing the difference between the simulated data and the experimental results. Using $\mathbf{F}$ as the objective function and shifting factors as features, the next step is to build up surrogate models and make predictions.

3.2.3 Adaptive optimizer

Considering the relatively high cost of simulation models, direct optimization that requires a large number of iterations is usually infeasible. An adaptive optimizer [91] has been developed to address such challenges: it combines surrogate modeling [88] and infill criterion [92] to refine the prediction adaptively while minimizing the sampling cost. Such adaptive strategy has been applied to sequentially optimize and design new materials with target properties [93, 94] .

In our work, an adaptive optimizer is applied to construct the surrogate model on the current observations and suggest the next optimal sampling points. In this manner, the initial surrogate model is constructed based on a relatively small number of initial DOE points, and an infill criterion sampling strategy is then applied to determine additional sampling points to augment the training data.

The surrogate model $y(\mathbf{x})$ built on all available data set $\mathbf{d}_{current}$ is represented by $y(\mathbf{x}|\mathbf{d}_{current})$. In our case, the input $\mathbf{x}$ is either a five-dimensional shifting factor data set for dielectric study or a two-dimensional shifting factor set for viscoelastic case. The available data set $\mathbf{d}_{current}$ is all the shifting factors available at the current step. The surrogate model of the objective function is built on all available shifting factors. The sequential sampling strategy determines the next sampling point based on the current model and certain infill criteria that helps to minimize sampling cost as well as to find the global optimum when dealing with highly non-linear response surfaces. After the new design points d_{new} are determined, the surrogate model is updated based on the new dataset $(\mathbf{d}_{current}, d_{new})$. The surrogate model and the optimal solution of the shifting factors keep evolving until a stopping requirement is satisfied.

The surrogate model is able to predict the difference between simulation and experimental data with different shifting factors as well as quantify the associated uncertainties of the prediction. The choice of surrogate model is case dependent, and in this work, the Gaussian Process (GP) [95] is chosen because of its capability to capture nonlinear response surfaces, as well as the flexibility for assessing prediction uncertainties, which is important information needed in the successive sampling procedure. A GP $y(\mathbf{x})$ is completely defined by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x},\mathbf{x}')$:

$$\begin{aligned} m(\mathbf{x}) &= \mathrm{E}[y(\mathbf{x})], \\ k(\mathbf{x},\mathbf{x}') &= \mathrm{E}[(y(\mathbf{x})-m(\mathbf{x}))(y(\mathbf{x}')-m(\mathbf{x}'))], \end{aligned} \tag{4}$$

where $\mathbf{x}$ and $\mathbf{x}'$ represents different sampling points in the design space, $\mathrm{E}(\bullet)$ is the expectation. The GP can be written as:

$$y(\mathbf{x}) \sim GP\ (m(\mathbf{x}), k(\mathbf{x},\mathbf{x}')) \tag{5}$$

For our study, the GP model can be written as:

$$F_k \sim GP\ (m(\mathbf{x}\,|\,\mathbf{d}_k), k(\mathbf{x}\,|\,\mathbf{d}_k, \mathbf{x}'\,|\,\mathbf{d}_k)) \tag{6}$$

where F_k is the difference for k-th iteration as derived in equation (3), and $\mathbf{d}_k$ is the available data in k-th iteration. For simplicity the mean function is chosen to be equal to zero and the covariance function is in the squared exponential form, also known as Kriging [88]: $k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\{-|\,w\cdot(\mathbf{x}-\mathbf{x}')|^2\}$, where σ is a constant and w is the roughness parameter of

GP, whose dimension is the same as $\mathbf{x}$. With these notations, the uncertainty of prediction $s^2(\mathbf{x})$ is also tractable, and the detailed mathematical expression can be found in [89].

In general, by applying the GP model, a fitted surrogate model that may have several local optima with input/output (shifting factor set/difference) relationship can be obtained. To find the optimal shifting factors, the simplest way is to search for the minimum difference based on the surrogate model. However, this may lead to a local minimum because the procedure does not acknowledge the associated uncertainty of the model. Simply choosing the minimum points places too much emphasis on the predictor (optimal mean prediction) while ignoring points that have high uncertainties (measured by the standard error of the predictor). Therefore, as we search the design space, we must balance this competition. The Expected Improvement (EI) [91, 96] is an important infill criterion for GP-based global optimization problem. EI is defined as the expected improvement at a design point x, conditioned on the current model y:

$$EI(\mathbf{x}) = E\left[(y_{\min} - y(\mathbf{x})) \mid y\right], \tag{7}$$

where $y_{\min}$ denotes the current minimum of the model. By expressing the right-hand side of the equation in integral form and applying integration by parts, EI of a GP model can be obtained:

$$EI(\mathbf{x}) = \left(y_{\min} - y(\mathbf{x})\right)\Phi\left(\frac{y_{\min} - y(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{y_{\min} - y(\mathbf{x})}{s(\mathbf{x})}\right), \tag{8}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard Gaussian cumulative distribution function and probability density. It can be proved that the expected improvement EI is large where the solution

has high uncertainty (large $s^2(\mathbf{x})$) and small mean prediction $y(\mathbf{x})$ [91] . The new sample point is selected as the maximizer of EI:

$$\mathbf{d_{new}} \in \arg\max_x EI(\mathbf{x}) \tag{9}$$

As illustrated in Figure 8, the new sample points, with their simulated properties from FEA, augment the initial training set to improve the accuracy of the surrogate model and the according prediction as expressed in Equation (10).

$$\begin{aligned} &\mathbf{d}_{k+1} = (\mathbf{d}_k, \mathbf{d}_{new}), \\ &F_{k+1} \sim GP\ (m(\mathbf{x} \mid \mathbf{d}_{k+1}), k(\mathbf{x} \mid \mathbf{d}_{k+1}, \mathbf{x}' \mid \mathbf{d}_{k+1})) \end{aligned} \tag{10}$$

Finally, as the objective is to minimize the difference between the simulation and experiments, the optimal solution predicted at each step is determined as the minimizer of objective function based on the current surrogate model.

$$\mathbf{x\_opt_k} = \arg\min_{\mathrm{x}}(F_k) \tag{11}$$

3.3. Results and Discussion

In this section, a representative dielectric data set (for a nanocomposite composed of 2 wt% bimodal anthracene-PGMA grafted silica in epoxy [38, 97]) is first selected to develop and test our methods. Detailed intermediate results are given to illustrate the methodology. Later, this algorithm is further tested on viscoelastic studies with the experiment samples collected from nanocomposite samples composed of 2 wt% Chloro-modified nanosilica in Polystryrene (PS)

nanocomposites [63]. We have TEM images and reconstructed microstructures for both the viscoelastic and dielectric samples as shown in Figure 11. Note that the microstructures are statistically equivalent to the actual samples and are created using an algorithm developed earlier [98].

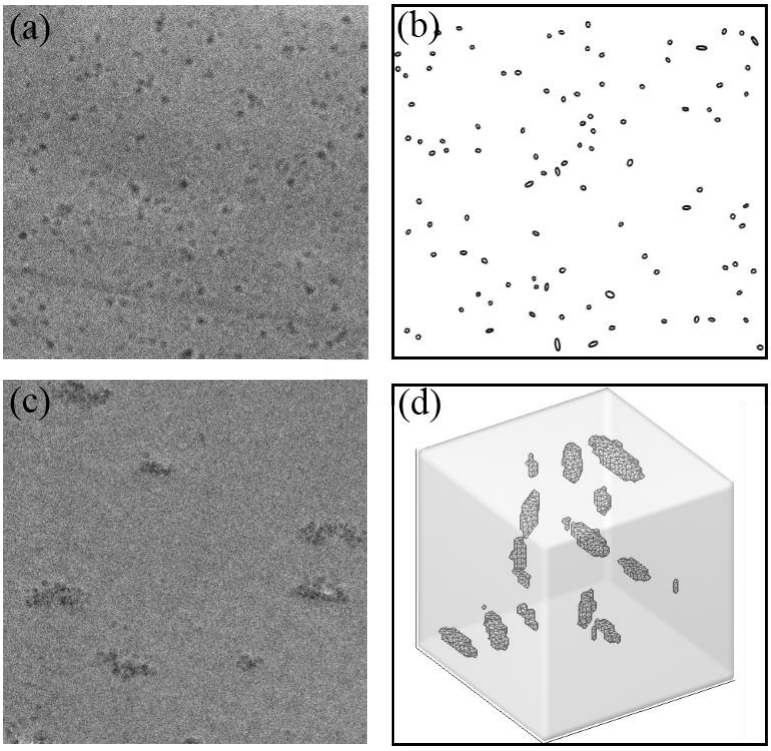

Figure 11. TEM images and the reconstructed microstructure used in FEA. (a) 2 wt% bimodal anthracene-PGMA grafted silica in epoxy TEM image, (b) reconstructed 2D microstructure of (a), (c) 2 wt% Chloro-modified nanosilica in PS TEM image, (d) reconstructed 3D microstructure of (c).

In our previous work [87], we determined manually fitted shift factors for a number of dielectric nanocomposites and using this prior study, we can select reasonable ranges within which the shift factors may vary (Table 1).

Table 1.dielectric shifting factor design space

Shifting factor	S_β	M_β	S_α	M_α	c
Range	0.4-1	1.5-3	0-0.2	0.7-2	0.3-2.5

For this dielectric study, 20 sets of shifting factors are generated as the initial data set within the empirical ranges through design of experiments (DOE) created by applying the optimal Latin hyper cube (OLHC) sampling technique [99]. The sample points are selected to cover the entire input variable space bounded by the given ranges (the blue points in Figure 13b-d illustrate the initial DOE for this example). Given the microstructure obtained from the TEM image for the chosen sample (Figure 11a,b), FEA simulations are performed over the 20 OLHC samples for shift factors to obtain the composite dielectrical response. The FEA model generates two outputs on each simulated sample: frequency-dependent storage and loss dielectric (or viscoelastic) curves. These results are then processed by calculating the difference between the simulated curves and the target experimental data using Equation (1). For the chosen dielectric data set, the real and imaginary differences, $\mathbf{f}^1$ and $\mathbf{f}^2$, characterized for the initial 20 shifting factor sets are plotted in Figure 12.

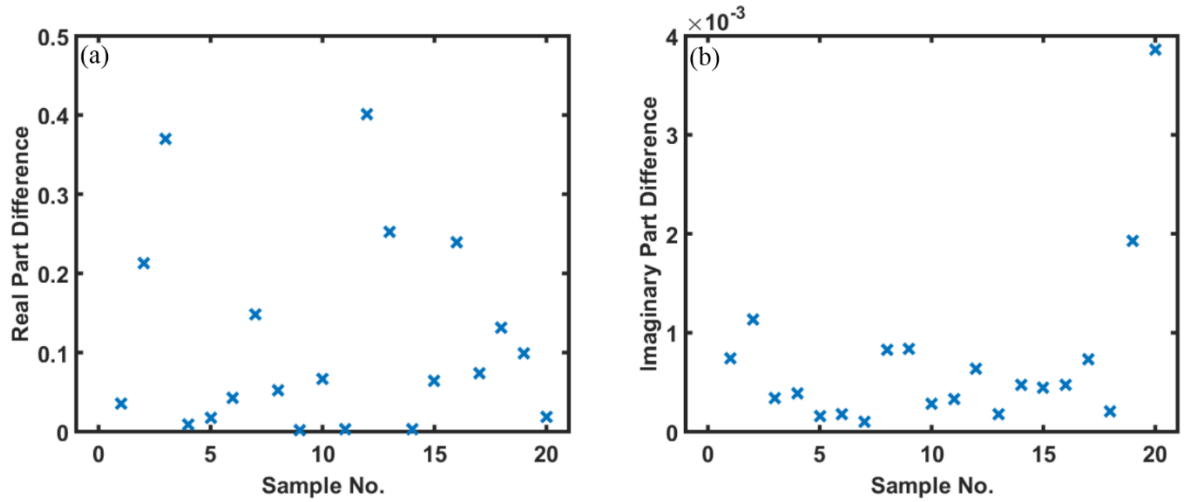

Figure 12. Differences between 20 DOE designed simulations and experiment data as applied to 2 wt% bimodal anthracene-PGMA grafted silica in epoxy sample. (a) Real part difference characterized by MSE;(b) Imaginary part difference characterized by MSE

It can be seen that the difference calculated from two parts have significantly different orders of magnitudes, which makes direct comparison difficult. Therefore, it is necessary to normalize the magnitudes using Equation (2) and formulate a multi-objective optimization problem applying Equation (3) such that minimizing the difference from both real and imaginary data simultaneously can be achieved. Below, an adaptive optimizer is applied to navigate the high dimensional design space and make accurate predictions.

In our study, our objective is to minimize the difference, for which the ideally optimal value of 0 indicates a perfect match between the simulated and experimental results. Practically, we find a threshold value C, indicating an acceptable fit between the experimental data and simulated results. By applying an adaptive optimizer, at each iteration, a new sampling point is given based on the feedback from the optimization to augment the training set and the iteration will cease after the output difference is less than the threshold:

$$\min(F_{\mathbf{k}+1}) \leq C, \tag{12}$$

where F_{k+1} is the surrogate model at k+1 step determined using Equation (10). The threshold *C* may vary from case to case while the *C* is set to 0.01 in our study. Additionally, to avoid infinite loop of the adaptive optimizer arising from bad data or parameter limits, the loop is set to terminate after this is no sign of decreasing difference after 20 iterations.

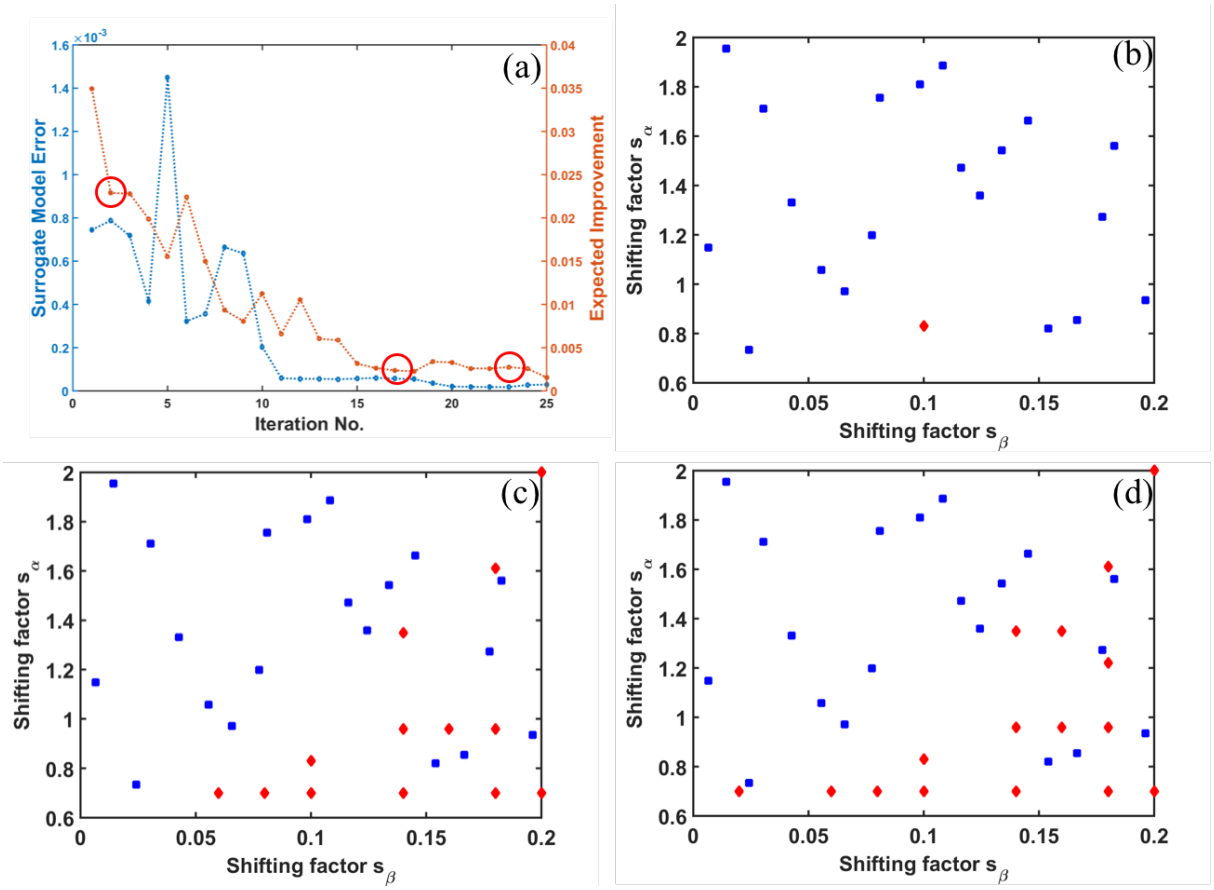

Figure 13 Evolution of the surrogate model with successive iterations and training data distribution for 2 wt% bimodal anthracene-PGMA grafted silica in epoxy dielectric sample; (a) Surrogate model error and Maximum EI (Equation (8)*) as a function of iterations, where the surrogate model error is the evaluated on a validation set with sample size 30; (b) Training sample distribution on iteration 2 (blue points are initially sampled from DOE while the red is sequentially added by Maximum EI); (c) Training sample distribution on iteration 17; (d) Training sample distribution on iteration 23*

In order to evaluate the prediction accuracy of the surrogate model at each iteration, a separate validation set with sample size 30 is generated by random sampling another 30 shifting factor sets within the design space and running simulations accordingly. Then the difference

between the simulations and the experiment for the validation set can be evaluated using Equation (3).The surrogate model error at each step is then evaluated by calculating the mean square error between the predicted difference from the surrogate model with that from the validation set. Figure 13 shows the evolution of the surrogate model accuracy as iteration progresses in the adaptive optimization process. In order to visualize the sample distribution during the optimization process, we choose two factors (S_α and S_β) and plot the distribution on iteration 2, 17, and 23 respectively (Figure 13 b,c,d). On iteration 1, only the initial 20 training points from OLHC (blue points in the Figure 13b) are used to make predictions. Then later, at each step, a new sampling point (red in Figure 13b) is added based on the maximum EI criterion. The newly added points are concentrated on the boundaries and the space with coarse initial sampling showing that the adaptive optimizer is able to automatically explore the entire design space with high uncertainties rather than search only in a local space which could result in a local minimum. It is noted that the error for each step keeps decreasing (though not monotonically, likely due to the model uncertainties) indicating an increase in model accuracy as new sampling points are added. Additionally, the dropping EI in the model also suggests the decreasing uncertainty of the surrogate models and the increasing model accuracy with the identified shift factors.

Figure 14 shows how our predicted optimal solution ($\mathbf{x\_opt_k}$ from Equation (11)) behaves as a function of iterations. The prediction accuracy at each step is evaluated by calculating the discrepancy between the simulated data using the predicted shifting factor from the adaptive optimizer and the given experimental properties using Equation (3). Figure 14a shows that the prediction accuracy of $\mathbf{x\_opt_k}$ at each step increases as a function of iterations. We also show the

comparison between the simulated result and the experimental data on iteration 1, 12, and 20, showing that the improvement of fitting quality. Iteration 1 predicts the well at some frequency ranges, but the difference is still much greater than our threshold *C=0.01*(shown as dashed line in Figure 14a). As the iteration proceeds, the decreasing difference value indicates better fitting (iteration 12) and it takes 20 iterations before the difference is less than the threshold. Ideally, the iteration stops at iteration 20 based on our stopping criteria but we plot more iterations in order to show the convergence of our adaptive optimization procedure.

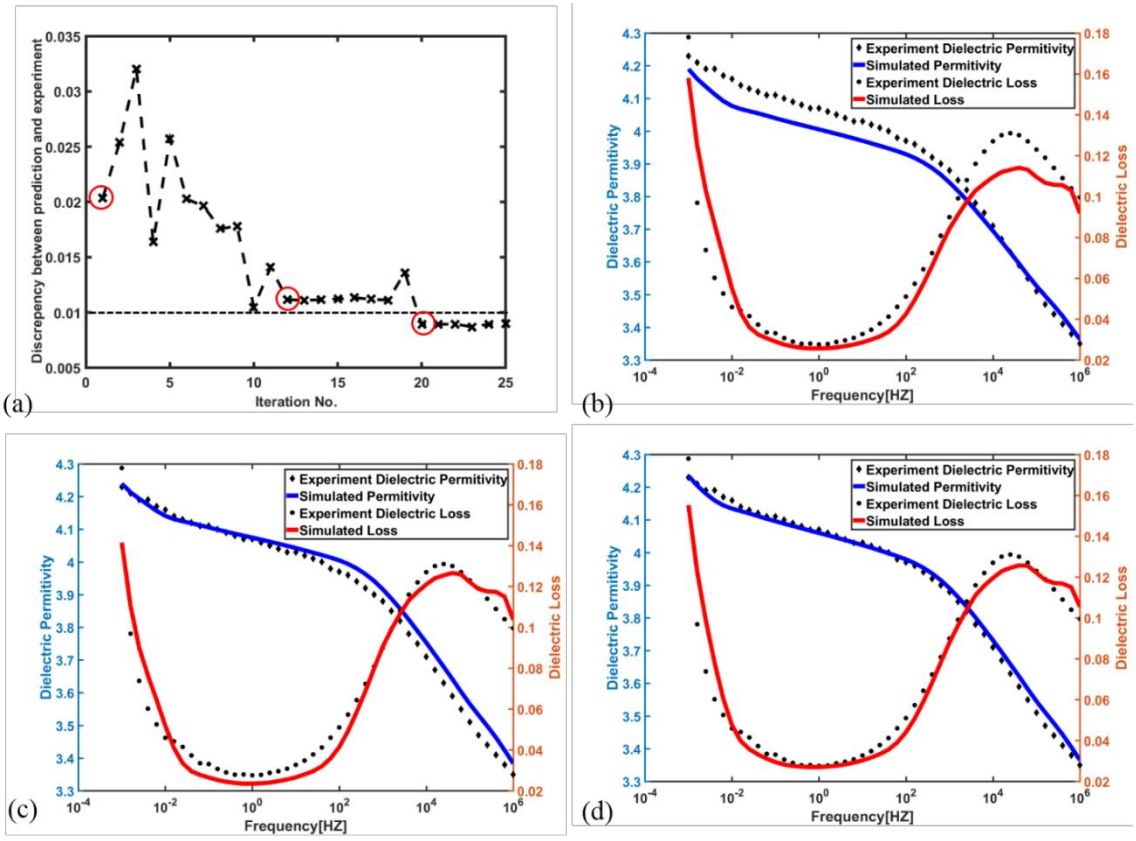

Figure 14 Evolving of optimal solution as a function of successive iterations; (a) The discrepancy between the optimal prediction and the experimental data as a function of iterations; (b) Comparison between experimental data and the simulated result on iteration 1; (c) Comparison on iteration 12; (d) Comparison on iteration 20;

Next, we want to compare different optimization strategies to illustrate that the adaptive GP model can reduce the computation cost and improve searching efficiency. Using the manual

fitting result as a reference, in addition to the adaptive GP model illustrated earlier, we also demonstrate the result of a one- stage GP and a random search model. The results from one-stage GP is generated by building the surrogate model based on all the sampling points in the data set and directly predict the minimal value of the objective function and the associated optimal solution at one time without sequential sampling. The result from the random search is generated by merely choosing the sample with minimal difference among all the randomly generated samples without doing any optimization. Figure 16 shows comparisons between these different algorithms by plotting the fitted curves and the experimental data in the same figure. All the searching algorithms are compared under the same prerequisite: the simulation cost is the same (i.e. the number of performed simulations is the same, in this comparison, the number of performed simulation is set as 40). To compare the performance of each method, we also calculate the discrepancy between the simulated data using the predicted shifting factor and the given experimental properties using Equation (3) as shown in Table 2. By comparing the fitting quality, we find that the performance of the adaptive GP is at least comparable or even better compared with the manual fitting quality.

Table 2 Comparisons of different searching methods by calculating the difference

Method	Discrepancy between the simulation and experiment
Adaptive GP	0.0089
Manual	0.0097
One-stage GP	0.1683
Random	0.1854

However, it is noted that the manual fitting requires guessing the optimal solution based on current fitting while the adaptive GP is able to automatically explore the design space and

determine the optimal solution iteratively. Additionally, the comparison between the adaptive GP and the one-stage GP illustrates the advantages of the sequential sampling approach. The surrogate model is more accurate when determining the next sampling points 'intelligently' through calculating the EI based on the current surrogate model than sampling all the points at one time. All optimization-evolved searching algorithms tested are better than the random search. This result indicates that the prediction accuracy of the adaptive GP model is the highest comparing with other searching strategies (random searching and one-state GP) if the computation cost is required to be the same. In other words, more simulations need to be performed for other models to generate comparable result as adaptive GP. These comparisons indicate that our adaptive GP model is able to accelerate the search process while maintaining a similar accuracy compared with the manual fitting procedure.

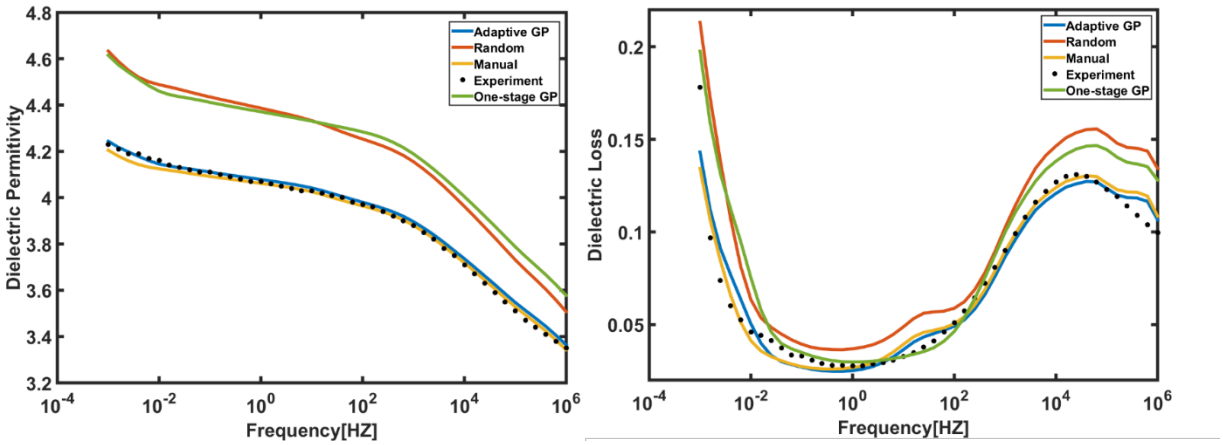

Figure 15 Performance comparison of different searching strategies (Adaptive GP, one-stage GP, Random, Manual)

By applying the adaptive GP model as described above, the optimal shifting factors can be determined. Since the shifting factors are descriptors for the interphase properties and completely define those properties, the interphase properties are thus determined. Figure 16shows the optimal interphase properties together with the matrix and composite properties. We can observe that the

interphase properties are significantly different from the neat polymer matrix. This change in properties of the interphase is expected due to changes in the mobility of polymer chains near the particles which cause changes in the local physical properties such as dielectric spectra of the polymer. This interphase regime shows a higher dielectric permittivity and loss than both the matrix and composite data. This interphase property agrees with the experimental data where the addition of functional groups on fillers enhances the relative permittivity. Additionally, it is noted that the magnitude of the optimal fitted properties is dependent upon the size of the interphase regime chosen. While here we use a fixed size for the interphase zone based on prior studies on filler-filler spacing and interphase in nanodielectrics [9], the automated optimization procedure developed here can allow a thorough exploration of such variables and their impacts.

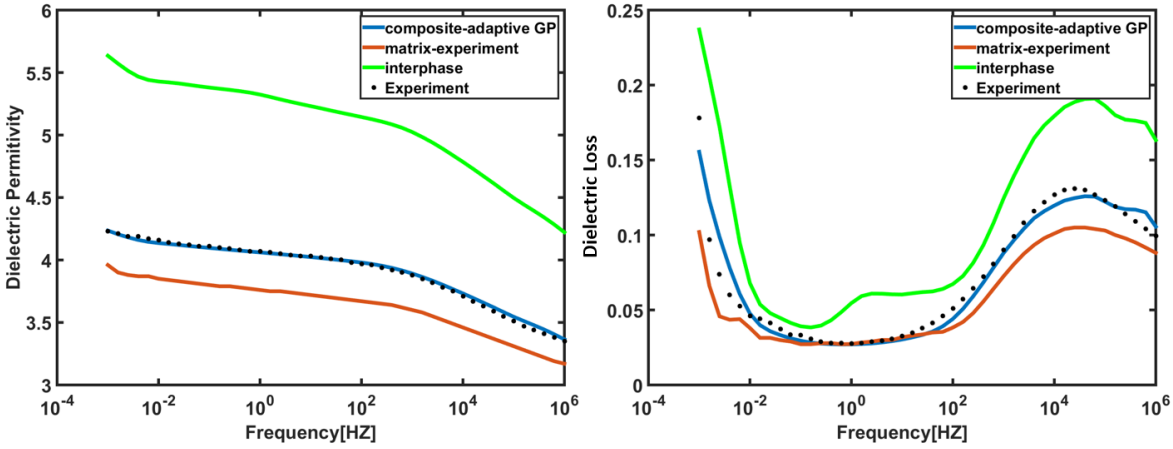

Figure 16. Comparison between simulated dielectric spectra using the shifting factors given from adaptive GP and experimental data for 2 wt% bimodal anthracene-PGMA grafted silica in epoxy

Figure 17 shows the comparison between experimental data collected from the chosen sample and the simulated viscoelastic response based on the predicted interphase properties from adaptive GP. With the predicted interphase shifting factors from adaptive GP, simulated response

is able to fit well with the experimental data. Similar to dielectric study above, the interphase properties are readily obtained directly from the optimal shifting factors. Compared with the dielectric study, it takes only three iterations for our adaptive GP model to obtain the optimal result while 20 iterations are required for the dielectric case. The two-dimensional space of the viscoelastic study allows the initial OLHC points to be nearly sufficient to build accurate surrogate models, while for the higher dimensional dielectric study, initial OLHC points is too sparse to construct accurate response surface and further iterations were required.

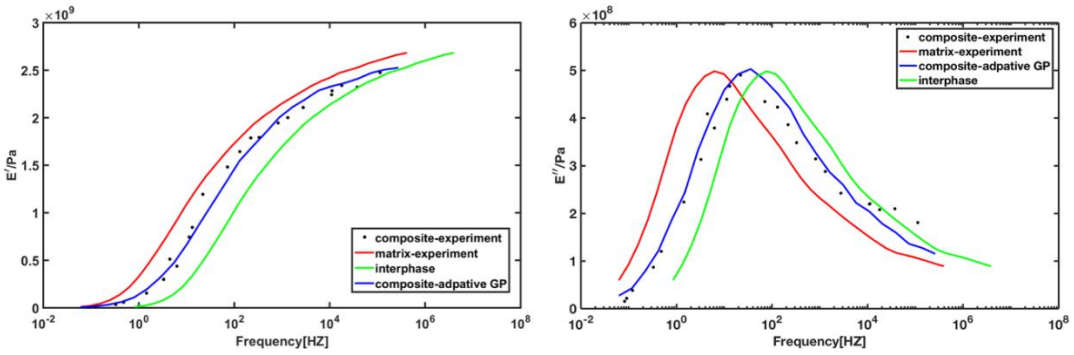

Figure 17. Comparison between simulated viscoelastic response and experimental data for 2wt% Chloro-modified nanosilica in PS

3.4. Conclusions

We have demonstrated a consistent and efficient approach for identifying the interphase properties in polymer nanocomposites by solving the inverse problem using adaptive optimization, improving prior work which was limited in accuracy and efficiency by the laborious manual iteration process. A multi-objective optimization problem is formulated through characterizing the difference from simulations and experimental data with the aim of minimizing the difference from real and imaginary parts simultaneously for frequency-dependent properties. The training data

consists of shifting factor sets as optimization variables and difference as objectives. The Gaussian Process (GP) surrogate model is used to build up the relationship between the features and objectives because of its flexibility for assessing uncertainties and capability to capture nonlinear response surfaces. By choosing candidate points based on the concept of maximum expected improvement (EI) over the search space, we have shown that the surrogate model evolves and the uncertainties in the model decrease by sequentially adding points at each iteration using the EI criterion.

Using the proposed approach and given experimental data of properties and microstructure, the interphase properties on dielectric and viscoelastic studies can be determined automatically. Results demonstrate that only tens of iterations are required for the method to identify the optimal shifting factors and interphase properties to achieve a good fit with the target experimental data. Comparisons are made among different searching algorithms to show the ability of our method to reduce the computational cost and improve the searching efficiency.

This method can be used as a generalized automated approach to determine the interphase properties in polymer nanocomposites. The framework is also flexible to be applied to other computational models to identify the interphase properties that are required. Additionally, the method is very efficient compared with the manual-fitting process and should facilitate the further investigation of a data-driven analysis where many hundreds of samples would be required. The future work may start from setting up a material interphase library through combining this method with *Nanomine*, which is a data-driven web-based platform for analysis and design of polymer nanocomposite systems under the material genome concept [63]. With sufficient data in the interphase library, relationship between the material constituents, microstructures, and the

interphase properties can be created through data mining. Such relationships have the potential to improve the understanding of material principles behind interphase properties and guide the design of new materials with desired functionality and performance.

Chapter 4. Mining structure-property relationship in polymer nanocomposites using multi-task convolutional neural network

4.1. Introduction

Data-driven methods have attracted more attention and become a hot-topic in material research after the proposing of material genome initiative. Combining material science with computer science and statistics, data-driven methods are aimed to expediate the material research and applications using past research data. In order to facilitate polymer nanocomposite data archiving, sharing and analysis, a data-driven web-based platform, NanoMine, has been developed for analysis and design of polymer nanocomposites. NanoMine consists both experimental and computational data on nanocomposite processing, structure and properties. In this paper, we present a novel deep learning approach that build the structure-property relation utilizing the data from NanoMine. Although a comprehensive study using experimental data is not feasible because of the limited data volume and variety in experimental conditions and material constituents, computational data allows demonstration of the approach and gives flexibility to sufficiently explore a wider range of structures. Taking advantages of microstructure reconstruction method and finite element simulations, we first explore qualitative relationships between microstructure descriptors (i.e. volume fraction and dispersion parameter) and mechanical properties. Then we present a novel deep learning approach that combines convolutional neural network with multi-task learning for building quantitative correlations between microstructures and property values. The performance of the model is compared with other state-of-the-art strategies including two-

point statistics and structure descriptor-based approaches. Lastly, the interpretation of the deep learning model is investigated to show that the model is able to capture physical understandings while learning.

Polymer nanocomposites includes organic matrix and nanoparticles with at least one dimension below 100nm [100]. By including a small amount of fillers to the matrix, experiments have shown that polymer nanocomposites are able to achieve a significant improvement in dielectrical, mechanical and optical properties compared with their parent matrix system [4-8]. The enhancement in properties comes from the large interphase region resulting from the strong chemical and geometrical interactions between the particle surface and polymer area nearby [9, 10].

To predict the properties of nanocomposites given structures, people have developed different types of numerical methods including continuum mechanics methods and multi-scale simulations [27, 63, 87]. A variety of micromechanical models such as Mori-Tanaka, Halpin-Tsai and the self-consistent scheme have been developed to predict the thermomechanical behavior of nanocomposites [50]. However, those analytical models are not sufficient to fully capture the dispersion state or the morphology information of the fillers although some of the models include structural parameters. Therefore, multi-scale simulations are often necessary. Finite element (FE) simulations are able to fully capture the structural information and accommodate non-homogenous material systems with explicit configurations of all material phases, which makes FE simulation a good candidate to analyze the behavior of nanocomposite. FE models have been developed to simulate the thermal and mechanical properties of polymer nanocomposites and investigate the impact of interphase [37, 49]. It has also been shown that the interphase properties in the FE models

can be represented by shifting factors based on the pure matrix properties for some of the material systems. The necessary shift for a given experimental sample can be determined automatically through Bayesian optimization [54].

In order to better understand the behavior of these materials and further design materials with target properties, researchers have proposed the paradigm of process-structure-property (PSP) linkage [101]. Data-driven approaches founded on PSP linkage have attracted more attention recently and become a hot topic in many areas of material research [61, 63, 102]. Data-driven approaches combine material science, computer science and statistics with the goal to expediate material discovery by utilizing past research data. To facilitate the development of data archiving, sharing and development of data-driven approaches, efforts have been made to create online databases for fast queries and reference. For example, Brinson group developed a polymer nanocomposite data resource called Nanomine, which allows fast data queries, visualization, sharing and analysis [63, 103]. Utilizing past research data and databases, data-driven approaches have been developed to model the P-S-P relationship. Hassinger et al. have developed quantitative relations between the nanocomposite processing parameters and the structural descriptors using a database of nanocomposites [63]. A new data-driven framework is proposed for designing and modeling of new material systems and structures, which includes design of experiments, computational models and machine learning methods. A common characteristic involved in almost all the data-driven approaches is that a relatively large pool of data is required so to apply machine learning methods can be applied to extract features from the data and build correlations.

Deep learning, one of the sub-field of machine learning, has dramatically improved the state-of-art in computer vision, natural language processing and many other fields including

material science [76]. Deep learning approaches provide an end-to-end framework addressing automated feature extraction for a large set of potential features. Using these approaches, it is not necessary to design explicit features, which is usually require in traditional machine learning approaches. Deep learning has been applied in material science specifically for the case where structural images are involved. Yang et al. applied deep convolutional network to model structure-property linkages for high-contrast elastic 3-D composite microstructures [102]. In [102, 104], deep neural networks are applied to reconstruct microstructure images and implemented to design material microstructure with desired properties.

In this paper, we investigate the structure-property linkage for polymer nanocomposites using a database of simulated data. It might be more appealing to build such linkage using experimental data from databases such as NanoMine, we found it extremely difficult to find systematic trend because of the scarcity of the data volume and the variety of material systems, processing conditions and even measurement methods coming from various papers and laboratories. On the other hand, simulated data gives flexibility to explore a wider range of structures and the data can be accumulated in a relatively short period. Therefore, in this paper, simulated data are applied to demonstrate the power of NanoMine database and similar study can be done in the future if a much larger set of experimental data is available. Although previous studies have demonstrated data-driven approaches to build structure-property linkages, a systematic study showing the impact of different factors (composition, dispersion, reconstruction method etc.) to different properties is still lacking. Additionally, most of the previous studies relied on traditional machine learning method using hand-crafted features. And the quality of the prediction model is mainly dependent on the quality of designed features from the material

expertise. And it is nontrivial to conduct feature selection from a big pool of microstructural features. In this study, by leveraging artificial generated microstructures and FE simulations, a sufficiently large database with different microstructures and associated properties can be obtained. The property of interest for this research is three mechanical properties of the bulk composites: $\tan\delta$ peak, glassy modulus and rubbery modulus. We aimed to investigate the effect of microstructure to multiple target properties in a more systematic and comprehensive way. The first part of the paper focuses on investigating the effect of microstructure reconstruction method, filler composition and dispersion, and interphase on the bulk properties. After that, we design and explore the application of a deep multi-task convolutional network to quantitatively predict multiple property values given a microstructure image. This is the first time a deep multi-task learning model is applied to build structure-property linkage for prediction of mechanical properties of polymer nanocomposite. The results from proposed method is further compared with other methods including traditional approaches using physical descriptors and two-point statistics.

4.2. FE Simulation and Analysis

4.2.1 Experimental data limitations

The properties of interests for this study are $\tan\delta$ peak, glassy modulus and rubbery modulus. This section focuses on qualitative understanding of structure-property relationship by plotting structure parameters with the property values. Based on physical understanding, the glassy and rubbery modulus should increase after more fillers adding to the system. On the other hand, the $\tan\delta$ peak should decrease at a higher loading because if replacement of damping material

(polymer matrix) with perfectly elastic particles (no damping). It is assumed that these relationships can be identified using a database with sufficient amount of data. Figure 18 shows the tan δ peak as a function of volume fraction for all the samples currently stored in NanoMine with volume fraction less than 10%. The plot shows no trend which contradicts our assumption. Given the condition that samples from NanoMine consist a wide range of material with different type of matrixes and fillers and also different processing, environment and even measurement methods, the data volume here is too small to restrict some of the parameters and study the effect of others and conduct a systematic study. Therefore, computation data is needed to give more flexibility to explore all different combinations and build a comprehensive linkage.

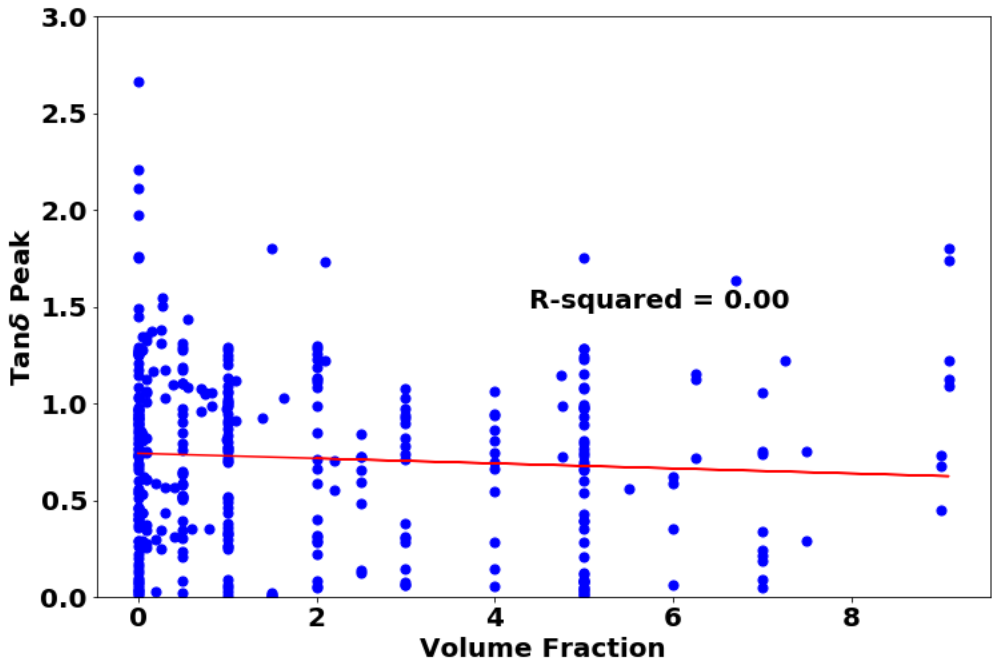

Figure 18. Plot shows tan δ peak as a function of volume fraction using the data from NanoMine.

4.2.2 Computation method

In order to demonstrate the application of data driven approach in correlating the structure-property relationship in polymer nanocomposites, we generated a set of simulated data using a combined of finite element (FE) method and microstructure reconstruction method.

The objective of this paper is to demonstrate data driven approaches in investigating the structure-property relationship. Due to the lack of good quality experimental data set for this purpose, we assume that the FE model could be accurate enough and well representative to simulate the associated property values. In the FE model, the interphase properties were assumed to be by shifting the polymer master curve two decades lower in the frequency domain assuming a stronger interaction between the nanoparticle and the polymer matrix. The FE configuration is shown in Figure 19. Polycarbonate (PC) is considered as the material property for the FEA and the frequency domain response of that material is measured by dynamic mechanical analysis (DMA). The silica nanoparticle in the model is assigned to be linear, elastic with young's modulus of 73MPa and Poisson's ratio of 0.3. For this study, we choose $\tan\delta$ peak, glassy modulus and rubbery modulus as properties of interest and correlate those with the structural inputs. The output from FEA is also frequency domain machinal response of the nanocomposite, based on which $\tan\delta$ peak, glassy modulus and rubbery modulus can be extracted.

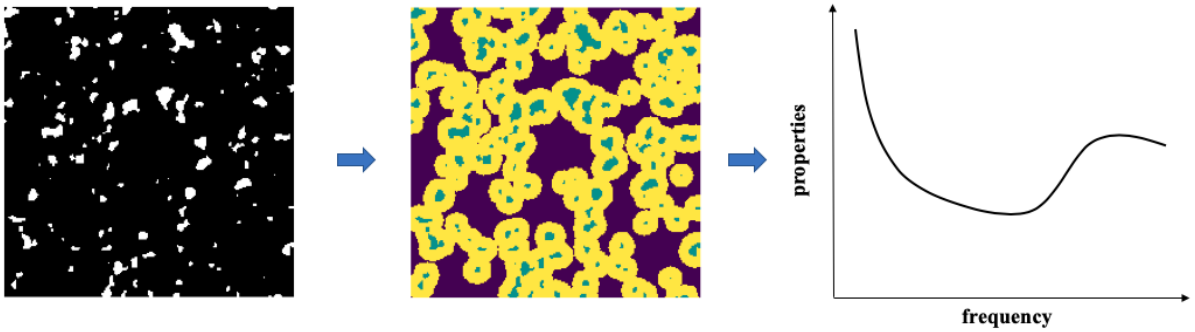

Figure 19. FEA configuration. Given a microstructure, an interphase layer is assigned assuming a fixed thickness and uniformed stronger interaction between the nanoparticle and the polymer matrix.

In this study, the structural inputs (referred as a representative volume element or RVE) for the FEA are generated using different microstructure characterization and reconstruction (MCR) methods, which are common applied in predictive material modeling [61, 104, 105]. Three different types of MCR methods are applied to generate RVEs for the FE model (i.e. uniform dispersion, descriptor-based method, and spectral density function (SDF)) and their impact on the structure-property relationship is later investigated. Microstructures with uniform dispersion are generated using the random sequential adsorption (RSA) algorithm such that the centroid of all existing particles have to exceed a minimum value and does not overlap with each other. Similar method as in this paper [37] is applied to generate the centers of spherical particles with same radius in the unit cell. Descriptor-based reconstruction is to generate statistically equivalent random digital microstructures by matching the prespecified structural descriptors (i.e. nearest neighbor distance, volume fraction etc.) through optimization [63]. By predefined a set of four descriptors, volume fraction (VF), number of clusters (n), average nearest cluster distance ($\bar{r}_d$),

aspect ratio (e_l), microstructure with different volume fraction and dispersion states can be obtained. The obtained microstructures have episode clusters which could overlap with each other compared with microstructures with uniform dispersion. Spectral density function (SDF) has been demonstrated to be sufficient to characterize complex heterogeneous microstructures and reconstruction can be done through phase recovery techniques [106]. We noticed that for our particle infilled nanocomposite system, the SDF of all microstructures approximately follows an exponential distribution with two parameters – shape variable α and scale variable θ. Our previous study has also shown that for our material system, the shape parameter α varies in a small range and has very little influence on SDF profile and the consequent microstructure [107]. However, the range of scale parameter have a wide range and could greatly impact the decay of SDF and the dispersion state of microstructure [106]. Therefore, for this study, a set of microstructures are generated by controlling the volume fraction (VF) and scale variable θ while the shape parameter α is fixed.

4.2.3 Impact of microstructure reconstruction method

Three representative microstructure reconstruction/generation methods are evaluated in terms of their impact to the property. For every reconstruction method, a hundred microstructures with different volume fraction (1% - 20%) and dispersion states are generated using Optimal Latin Hypercube sampling (OLHC) [99]. Microstructures obtained from different method are shown in Figure 20 . Here, we merely compare the impact of reconstruction method and there is no interphase layer added for the computation. The impact of interphase will be discussed latter. Microstructures with uniform dispersion have spherical particles, which are uniformly distributed

to the matrix and do not overlap with each other. Physical descriptor reconstructed microstructures have elliptical clusters with different aspect ratio and those clusters could overlap and form a bigger cluster. Microstructures reconstructed using SDF have irregular shape clusters. All the three microstructure types are realistic for different types of composite materials as shown in Figure 21. The simulated properties versus volume fraction are shown in Figure 22. The results show that for all the microstructures, the $\tan\delta$ peak decreases monotonically as a function of volume fraction while the glassy and rubbery modulus increases monotonically. This observation matches the physical understanding and result in a previous work [86]. $\tan\delta$ peak is a measure of material damping and the reduction of the magnitude coming from the replacement of polymeric (damping) material with perfectly elastic participle (no damping). It is also noted that as the microstructure becomes more complex (from uniform dispersion to SDF), the r-squared values decrease, indicating a less monotonical trend for all the three property values.

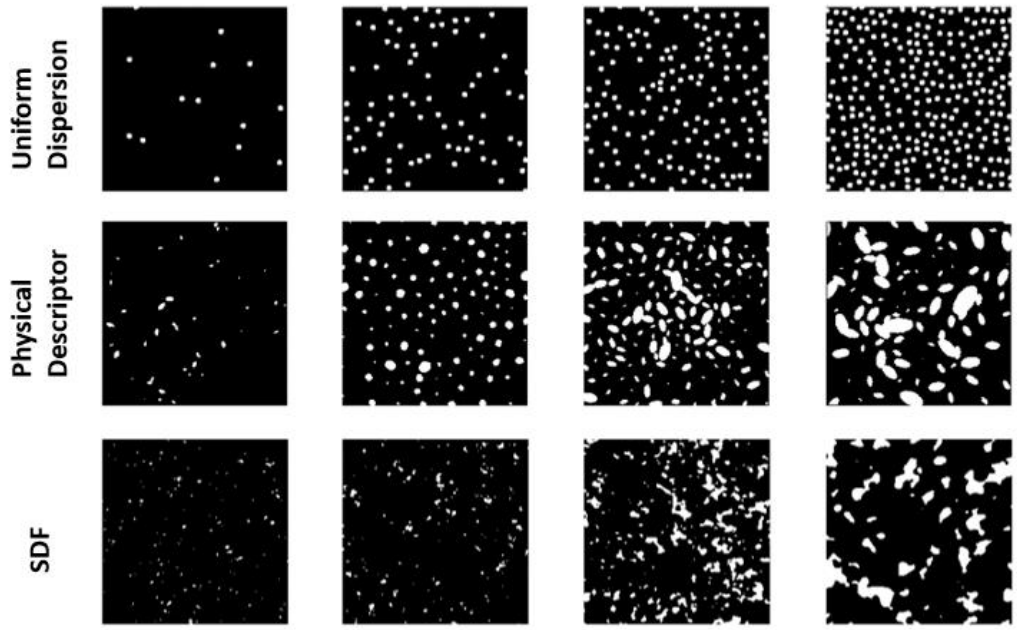

Figure 20. Microstructures generated using different methods: first row, uniform dispersion; second row, physical descriptor; third row, spectral density function;

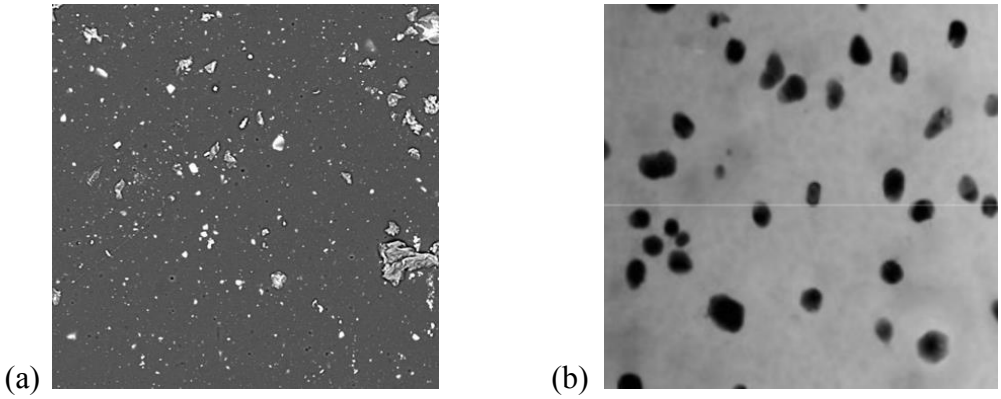

Figure 21. Experimental images showing different types of microstructures. (a) SEM images of octyl-modified silica in PMMA. (b) TEM images of Ag/C core/shell fillers in the epoxy [108].

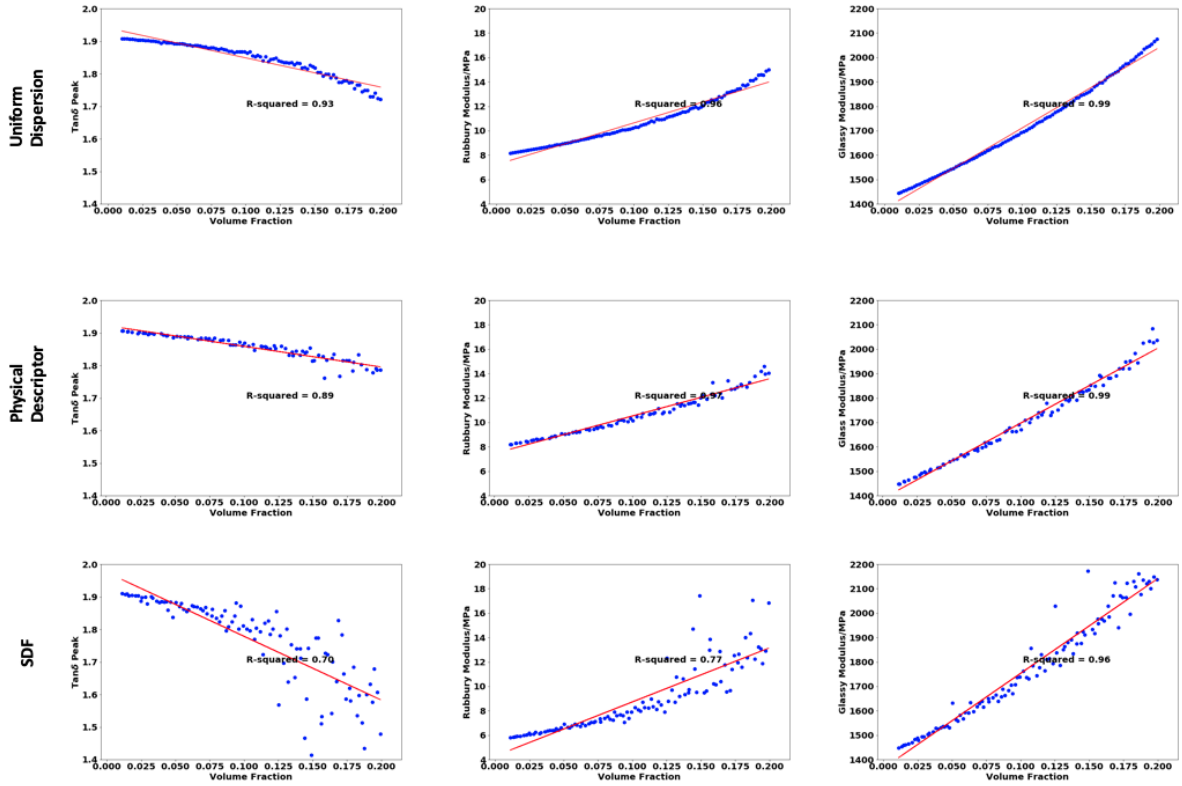

Figure 22. Comparison of simulations using different type of microstructures. For each type of microstructure, three individual property value (tan δ peak, glassy modulus and rubbery modulus) is plot against volume fraction.

4.2.4 Impact of interphase

Previous studies have shown that the interphase region plays an import role in the bulk composite properties [37, 49, 54, 63]. The interphase properties can be represented by shift factors related to that of the matrix. In this study, we assume a fixed thickness of interphase whose properties are determined by shifting the master curve of the matrix two decades lower in the frequency domain as shown in Figure 23. Although in this study, for simplicity, the interphase properties is assumed to be uniform, gradient interphase properties can be considered in the future

as well based on a new gradient interphase representation method in the FE model coming from observations from the local measurement of interphase [49].

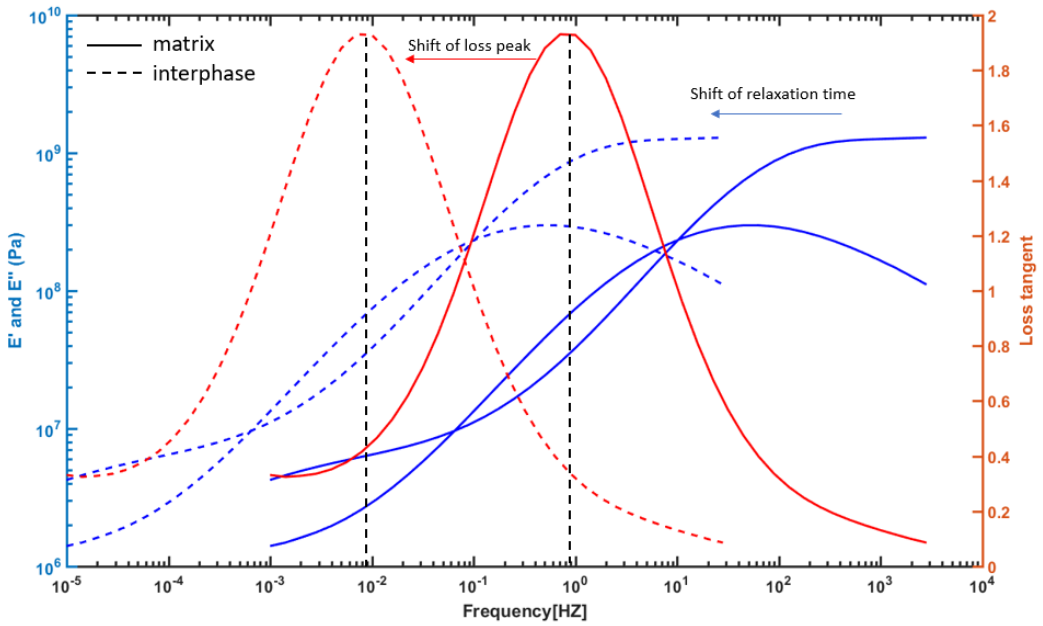

Figure 23. Relationship between matrix and interphase properties. The interphase properties are determined by shifting the master curve of the matrix two decades lower in the frequency domain.

FE simulations are run using the same set of microstructures using SDF approach as in 4.2.3 with interphase region. The results are shown in Figure 24. The property enhancement from the interphase layer exhibits an unbalanced reinforcement for the rubbery and glassy modulus, where the reinforcement of the rubbery modulus is more significant than that of the glassy modulus. While the variance remains similar, the glassy modulus exhibits almost the same range of increase with volume fraction as the no-interphase results while the rubbery modulus exhibits a much larger increase in magnitude compared to the no-interphase case. More significantly, the results for the $\tan\delta$ peak magnitude are dramatically different: $\tan\delta$ is nearly random with no discernable decrease with volume fraction, in contrast to the monotonic decrease for the simulations with no interphase. The results for the simulations with interphase for $\tan\delta$ are quite similar to the dataset

from NanoMine for a wide range of experimental data (Figure 18). This result can help explain the findings in Figure 18 from NanoMine: since the data from experimental samples are wide ranging and in many cases the samples will contain an interphase of altered properties near the filler, the lack of trend in $\tan\delta$ can be expected over that broad data set.

To better understand the results for the $\tan\delta$ trends, the volume fraction and dispersion parameter are restricted individually in different trials. The dispersion state of microstructures is controlled by the factor theta, which varies from [0.1,1]. Figure 25 shows the $\tan\delta$ peak as a function of volume fraction for data under similar dispersion, where the SDF parameter theta is restricted to different ranges ([0.1,0.3],[0.2,0.4],[0.6,0.8]) respectively. With constraint on the dispersion, the $\tan\delta$ peak exhibits a decreasing trend as a function of volume fraction at lower θ ranges ([0.1,0.3] and [0.2,0.4]), similar to the trend in the no-interphase calculations in 4.2.3. On the other hand, at a higher θ ranges, $\tan\delta$ peak shows an opposite trend: increasing as a function of volume fraction. This can be explained by the fact that the behavior of the nanocomposite is gradually changing from being dominated by the matrix to dominated by the interphase. Interphase volume fraction can increase both with dispersion of the fillers in the matrix and a larger filler loading. However, in the case of poor dispersions, interphase layers will overlap, leading to minimal change in interphase fraction with increasing filler loading. Therefore, at lower dispersion levels, the $\tan\delta$ peak value decreases with filler loading because the bulk property of the composite is controlled by matrix and filler and the replacement of polymeric (damping) material with perfectly elastic participle (no damping) decreases overall dissipation. At higher dispersion levels, the $\tan\delta$ peak value increases because the property of the composite is dominated by the interphase as the interphase volume fraction exceeds a percolation threshold [37]. This result

implies that for composites with interphase, both the loading and dispersion play a critical role in the determination of $\tan\delta$ peak values.

We are also able to investigate the converse situation where the volume fraction is restricted to different small ranges ([1%,5%],[5%,10%],[10%,15%]) respectively and the effect of dispersion is varied. Figure 26 shows the $\tan\delta$ peak as a function of dispersion for data under similar volume fraction. Figure 26a shows that at a lower volume fraction range ([0.01,0.05]), the $\tan\delta$ peak decreases as the dispersion increases. This suggests that at this lower loading, despite microstructures with different dispersions, the volume fraction of interphase does not exceed the interphase percolation threshold and the relaxation behavior of the composite is still being dominated by the matrix, with the peak located at the PC matrix frequency location (see Figure 23). As a result, as the dispersion improves, larger interphase volume reduces the tan d magnitude at the matrix peak frequency location (broadening the peak by shifting some magnitude toward the interphase peak frequency location) and results in lower $\tan\delta$ peak values. On the other hand, at higher volume fraction range (in Figure 26c, [0.10,0.15]), the $\tan\delta$ peak increases as the dispersion improves. This result can be explained by that fact that the property of the composite is dominated by the interphase as the interphase volume fraction exceeds the percolation thresholds at this higher loading and therefore more interphase leads to increasing the composite tan d peak which for the percolated cases is located near the interphase tan d peak frequency location. Figure 26b shows the $\tan\delta$ peak first decreases and then increases as a function of dispersion, which provides evidence of the transition of composite property from being dominated by the matrix to being dominate by the interphase.

Based on two controlled experiments, overall, adding interphase to the system greatly impacts the tan δ peak value and the value is dependent on both the loading condition (volume fraction) and the dispersion state (theta). These results also further clarify the no-trend result in the experimental data from NanoMine: these data sets contain microstructures with very different dispersions conditions, which must be accounted for to understand the damping response of the composite.

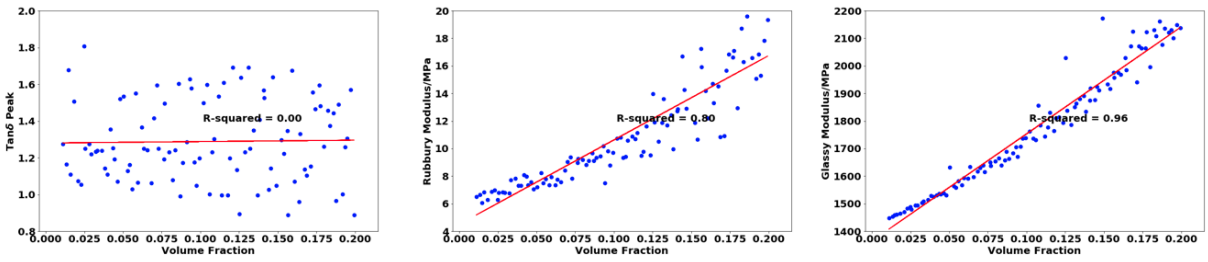

Figure 24. Results for simulations with interphase. The property shows tan δ peak, glassy modulus and rubbery modulus as a function of volume fraction.

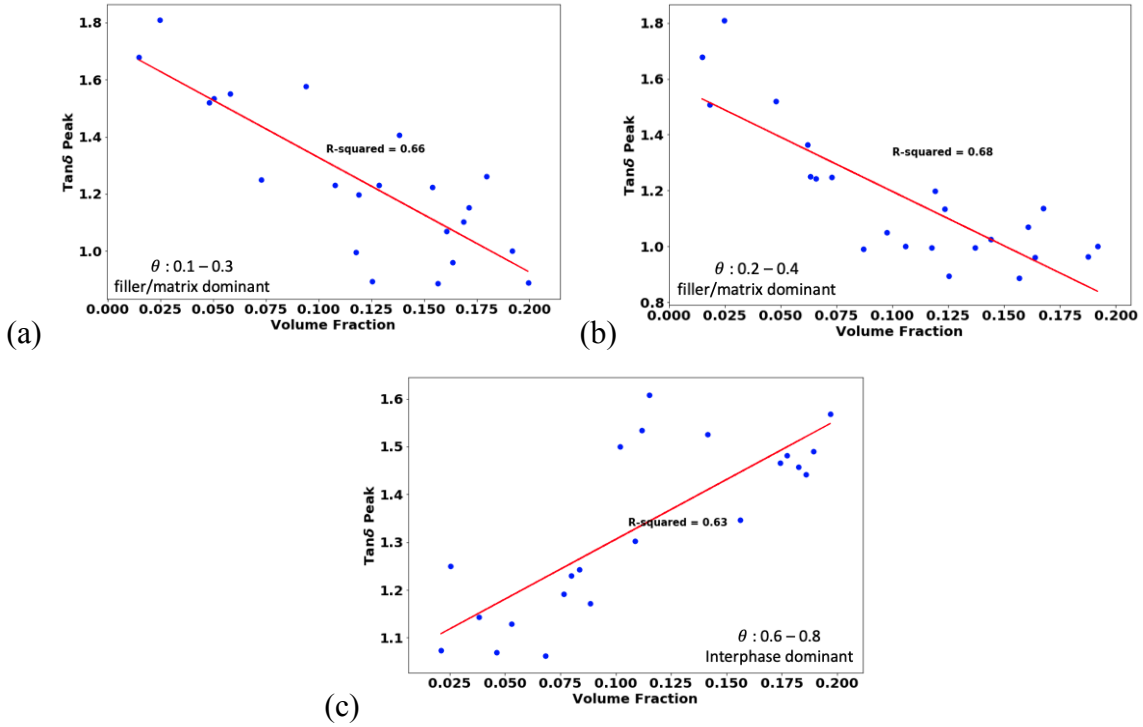

Figure 25. Tan δ peak as a function of volume fraction for controlled data sets with similar dispersion levels by restricting θ to (a) [0.1,0.3]; (b) [0.2,0.4]; ((c) [0.6,0.8]

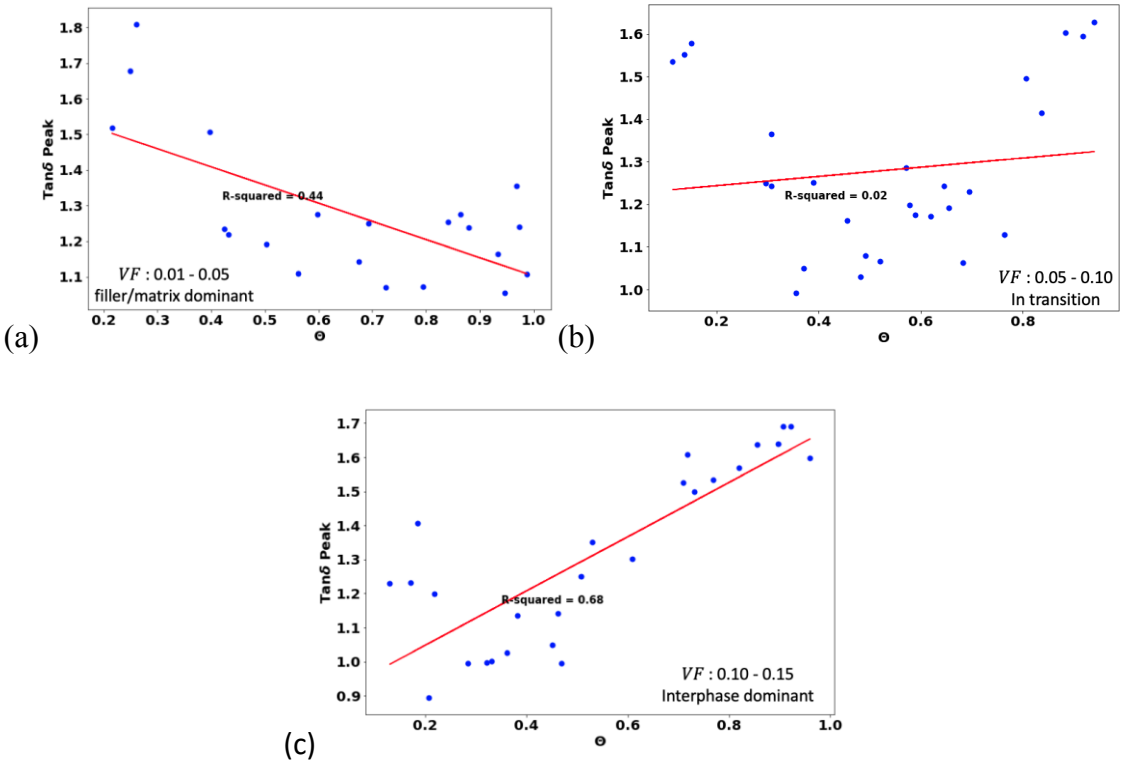

Figure 26. Tan δ peak as a function of dispersion θ for controlled data sets with similar volume fractions by restricting VF to (a) [0.01,0.05]; (b) [0.05,0.10]; (c) [0.10,0.15]

4.2.5 Deep learning model development

In earlier sections, we have qualitatively shown that the mechanical properties of the composites vary according to microstructures with different loading and dispersion. The next task in this paper is to quantitatively predict the properties of interests (tan δ peak, glassy modulus and rubbery modulus) given a microstructure. The objective is to predict continuous values given a 2D image, which can be considered as a regression problem using features from a 2D matrix. There are two common strategies that can be applied to build this regression model: one is to use hand-

crafted features (usually geometrical descriptors) that could sufficiently capture the characteristics of a given microstructure; the other is to use feature-engineering free method such as deep learning. The quality of the prediction using hand-crafted features is heavily dependent on the quality of designed feature from material experts and these features are usually not transferable from one material system to the other. On the other hand, compared with traditional machine learning approach using hand-crafted features, deep learning method is considered as a feature engineering free method that is able to automatically learn critical features during the training process, and have shown significant improvement in learning ability, generalization and transferability.

Convolutional neural network (CNN) is a deep learning approach and has been widely applied for computer vision related task, such as image classification and object recognition. In those tasks, CNN outperforms conventional method to a large extent due to its capability of extracting high-level abstractions of inputs through a series combination of non-linear transformations. A common CNN may include three different type of layers: convolution layer, pooling layer and fully connected layer. The objective of convolution layer is to extract critical feature maps from the input images through applying filters with different number and size. The values in those filters are learnt from available data. Pooling layer is often applied after convolution layers in the aim to reduce the dimension of the feature maps. Different down sampling strategies can be taken including max-pooling, average pooling and L2-norm pooling. By combining a series of convolution and pooling layers, a series of feature maps can be obtained and then fed to the fully connected layers for prediction of a class or a single value depending on the purpose of the task. The convolution and pooling layers are regarded as the feature extractor while the fully connected layers act as a non-linear regression model using the features from the first part. The

nonlinearity in the network is introduced through applying activation functions, such as ReLU, sigmoid, tanh.

Additionally, multiple outputs ($\tan\delta$ peak, glassy modulus and rubbery modulus) are required for our problem. In order to predict multiple values given a microstructure images, two strategies can be taken: one is to train separate machine learning models for specific tasks (i.e. for our purpose, three separate machine learning models are required to predict three different property values); the other is to develop a single machine learning model that solves multiple learning tasks at the same time, which is also called multi-task learning (MTL). Figure 27 shows two common ways to perform multi-task learning in deep neural networks: hard or soft parameter sharing [109]. The most common strategy is MTL is the hard parameter sharing, where several hidden layers are shared across all tasks while task-specific layers are applied after to predict the value for different objectives. It has been shown that hard parameter sharing could be able to reduce overfitting by including shared parameters. This is because the model has to find a hidden representations or features that captures all the tasks, which improves model generalization and reduces the chances of overfitting. For soft parameter sharing, each task has its own parameters. But the parameters for different tasks is constrained to encourage the parameters to be similar. Different constrains are usually applied such as L2 distance for regularization and the trace form. MTL has also been demonstrated to be extremely helpful if tasks share some similarities and the data volume is relatively small. MTL appears to be a good candidate for our purpose as the three learning tasks share significant commonalities, which are to predict mechanical responses given a set of microstructure images. Additionally, because of the expensive cost of FE simulations, it is time-

consuming to generate a very large volume of data. By utilizing MTL, it is aimed to implicitly augment the data, reduce the chances of overfitting and improve the model generalization.

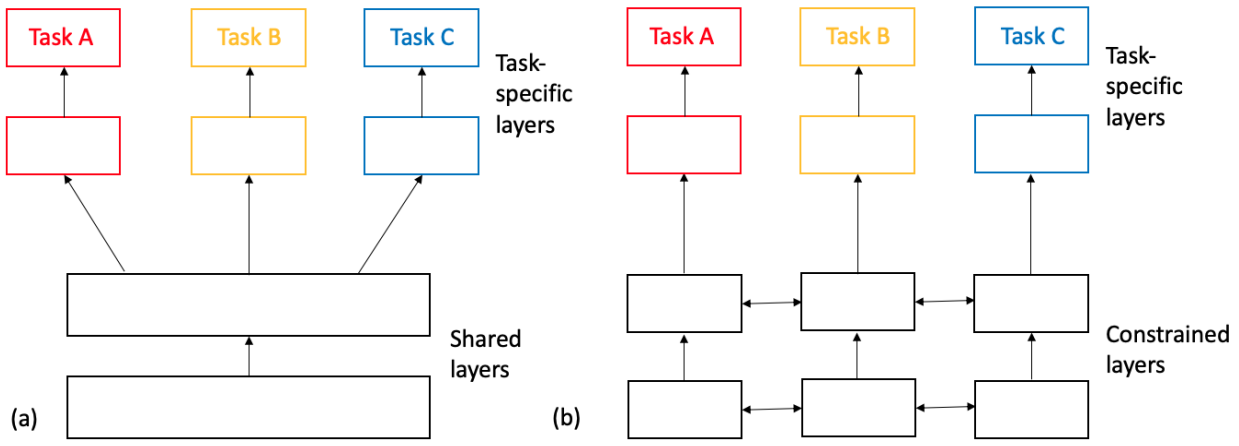

Figure 27 Architectures of two types of MTL models. (a) hard parameter sharing by applying hidden layers that shared across different tasks; (b) soft parameter sharing by having constrained individual parameters for different tasks to encourage similar parameters.

Taking advantages of the MTL and CNN, in order to properly capture and extract high-level microstructure features and build the linkage of structure to multiple property values simultaneously, a multitask CNN model is proposed. The overall architecture of the model is shown in Figure 28 and the detailed configurations and parameters of each layer is shown in Table 3. The input for this model is a 256*256 two phase binary image with zeros and ones, where zero represents the matrix and one represents the filler. Even though the FE model includes interphase layers in the calculation, the input for the deep learning model only consists two phases, filler and matrix. The image is pre-processed by replacing zero with -0.5 and 1 with 0.5, which has been found useful to improve the performance of convolution layers [102, 110]. The model consists a number of shared convolution and pooling layers and three sets of tasks specific layers including

convolution and fully connected layers. The shared layers, including convolution and pooling layers with different sizes, aim to extract the critical features from the images for determination of mechanical properties. Taking the features from shared layers, the task specific layers are designed to further extract task specific features and train different regressors for different outputs. The underlying reason for designing this architecture is that the extracted high-level features from convolutional layers should be shared across three highly related tasks while the weights for the later task specific layers are updated separately such that three different regressors are trained to predict different property values. The loss function for the model is formulated as the sum of mean absolute percentage error (MAPE) across three tasks:

$$\text{loss} = \sum_{j\in\{T,G,R\}} \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\overline{y_j^i} - y_j^i}{y_j^i} \right| \times 100\% \tag{13}$$

where j represents the type of task, for our case it only has three values: $\tan\delta$ peak (T), glassy modulus (G) and rubbery modulus (R). $\overline{y_j^i}$ represents the i-th predicted value for task j and y_j^i represents the true value and N is the total size of training data. The training of this model is achieved by minimizing the loss function through back propagation and optimization. Additionally, in order to prevent over-fitting, common methods in deep learning including batch normalization and dropout is applied.

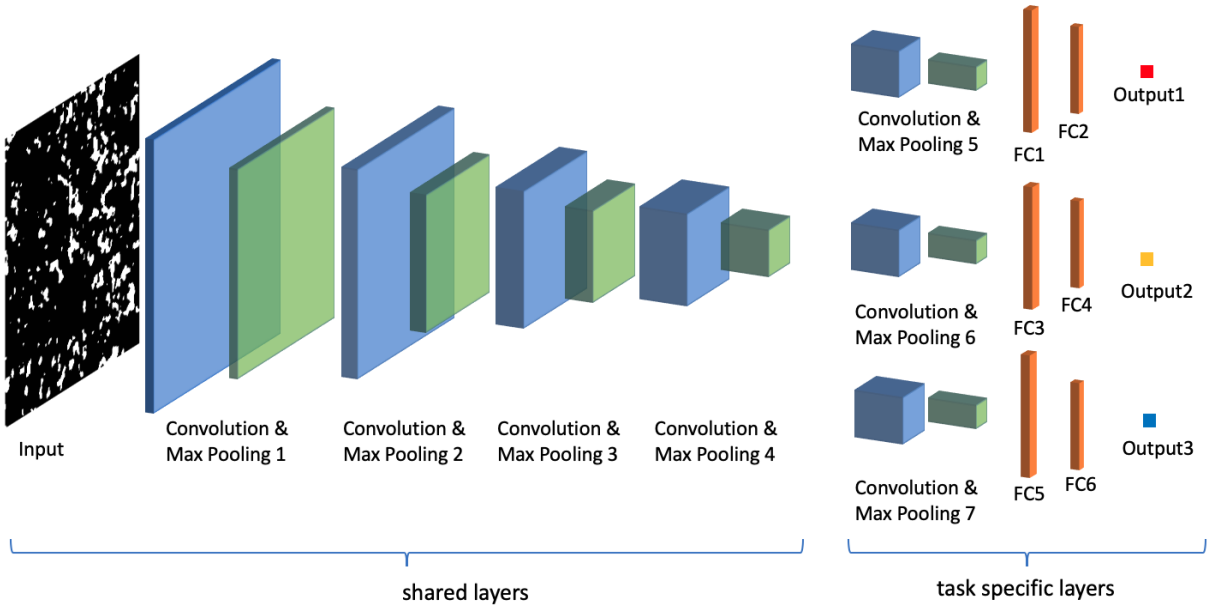

Figure 28 Architecture of the proposed multi-task deep CNN model. The MTL is achieved through hard parameter sharing. The input image is first feed to a series of shared convolution and pooling layers to extract the high level shared structural features for different tasks. Following that, three sets of task specific layers, including one more convolution and pooling layers and two fully connected (FC) layers are applied to predict different output.

Table 3. Model configuration of the proposed multi-task deep CNN model. For each conv block, 3×3 convolution and ReLU activation with back normalization (l2 norm rate = 0.0005) is applied. The pooling size is 2×2. After all the convolution and pooling process, the weights are flattened and feed to two fully connected (FC) layers.

Layer	Type	Dimension
Input	Shared	$256 \times 256 \times 1$
Conv and Pooling 1	Shared	$128 \times 128 \times 16$
Conv and Pooling 2	Shared	$64 \times 64 \times 32$
Conv and Pooling 3	Shared	$32 \times 32 \times 64$

Conv and Pooling 4	Shared	$16 \times 16 \times 128$
Conv and Pooling 5/6/7	Task Specific	$8 \times 8 \times 128$
FC 1/3/5	Task Specific	$1 \times 1 \times 512$
FC 2/4/6	Task Specific	$1 \times 1 \times 256$
Output 1/2/3	Task Specific	$1 \times 1 \times 1$

4.2.6 Datasets

In order to demonstrate the performance of proposed method, a data set with total size of 11000 is generated using generated microstructures and FEA model. Even though it might be more appealing if the structure property connection is studied using experimental data, a comprehensive prediction requires a large amount of training data that is not readily available for this study and not even available in the NanoMine databases. On the other hand, simulated data gives flexibility to explore a wider range of structures and the data can be accumulated in a relatively short period. Therefore, in this study, we take advantages of the simulated data and similar study can be done if sufficient experimental data is available.

Each microstructure, with dimension 256*256 pixels, is generated from SDF and considered as an RVE for the FEA. As illustrated in previous section, the volume fraction and the dispersion are controlled by the volume fraction (VF) and scale variable θ respectively. By sampling 11000 combinations of VF and θ, 11000 RVEs with different volume fraction and

dispersion are obtained. FEA is run for each RVE using Abaqus. Polycarbonate is chosen as the matrix material, whose master curve is measured through DMA. The filler is silica nanoparticle with a linear elastic young's modulus of 73MPa and Poisson's ratio of 0.3. The interphase layer has a thickness of ten pixels, whose properties are determined by shifting the master curve of the matrix two decades lower in the frequency domain. The FEA outputs the frequency response of mechanical properties of the nanocomposite, from which three properties of interest $\tan\delta$ peak, glassy modulus and rubbery modulus can be extracted. Figure 29 shows the property values as a function of volume fraction for those 11000 data. Compared with the previous results shown in Figure 24 using only 100 microstructures, the trend for $\tan\delta$ peak and glassy modulus is similar while the glassy modulus shows more variance. This is because 100 microstructures may not be sufficient to capture the whole space of different dispersions on glassy modulus.

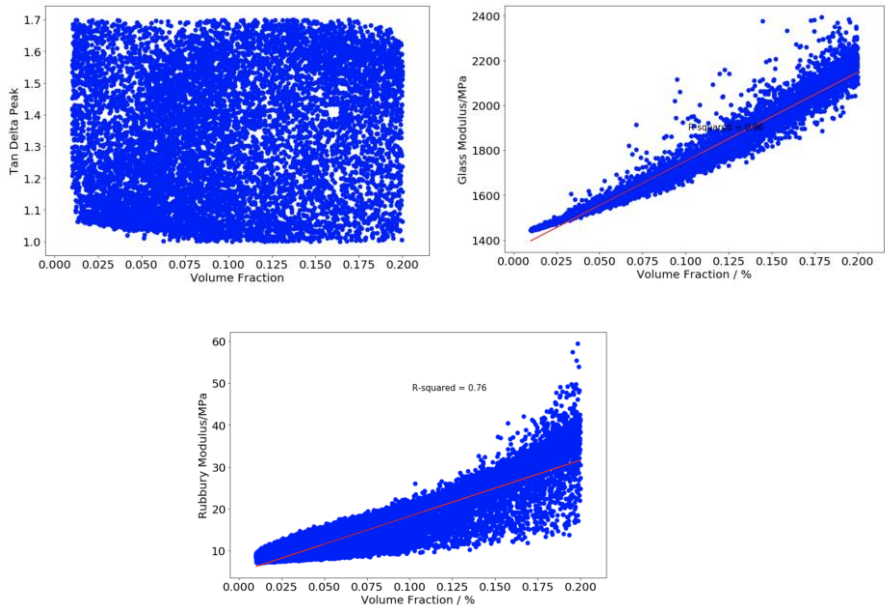

Figure 29. Simulation data for the deep learning model. The property shows $\tan\delta$ peak, glassy modulus and rubbery modulus as a function of volume fraction.

The 11000 simulated data is further divided into a training, validation and testing set with a ratio of 7:1.5:1.5: 7700 data are used to fit the deep learning model and 1650 is used to tune the hyperparameters in the model and the rest of data is reserved as testing data to evaluate the accuracy of the model. The accuracy of the model is evaluated using mean absolute percentage error (MAPE) which is given by:

$$e = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\overline{y^i}-y^i}{y^i}\right| \times 100\% \tag{14}$$

where $\overline{y^i}$ is the predicted value from the deep learning model and y^i is the true value from the FE model.

The FE model is run using Abaqus, a widely applied commercial software for FE simulations. The simulations are run on a work station with 192GB RAM and 16 core Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz. To build the deep learning model, Python 2.7 and Keras with Tensorflow backend is adopted. The model is trained on a work station with a NVIDIA Quadro P5000 GPU with 16GB GPU RAM and 20 core Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz with 192GB RAM.

4.2.7 Model performance

The performance of the model is benchmarked on two other common strategies for predicting structure-property linkage. Figure 30 shows different strategies to predict the properties given a microstructure image. Other than the deep learning based approach, one of the traditional methods is to build correlations using hand crafted features such as volume fraction, nearest neighbor distance, aspect ratio etc. [61, 111]. Those features are usually designed by domain

experts with the aim to capture the composition, dispersion as well as the geometric information of the microstructure. The structural features are then ready to be applied to fit a machine learning model to predict the property values. Another strategy to build the structure-property linkage is to compute two-point correlation functions from a microstructure and applied as features for the regression-based models [65, 70, 105]. Additionally, dimension reduction technique such as PCA is often employed on the two-point statistics because of the extremely high dimensionality. In order to evaluate the robustness and stability of the model, for every method, the model is trained and evaluated for ten times on different training testing splits. Additionally, to compare the performance of different methods, different models are given the same set of training, validation and testing set. The result is shown in Table 4.

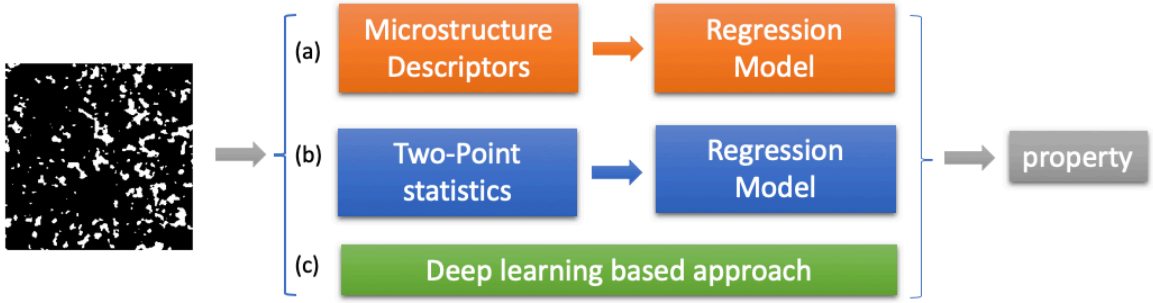

Figure 30. Different types of methods to predict structure-property relationship. (a) Geometric descriptor-based approach (i.e. application of hand-crafted geometric features such as volume fraction, aspect ratio etc. to a regression model)); (b) Two-point statistic methods (using two-point statistics as features for the regression model); (c) Deep learning-based approach (feature engineering free).

Table 4 Result comparison for different methods. The value shows the mean of MAPE together with the standard deviation across ten trials.

Method	Glass modulus	Rubbery modulus	$tan\,\delta$ peak
Two-point Statistics + PCA	$1.36\% \pm 0.03\%$	$5.75\% \pm 0.14\%$	$8.98\% \pm 0.19\%$
Geometrical Descriptors	$1.24\% \pm 0.08\%$	$4.74\% \pm 0.08\%$	$4.46\% \pm 0.10\%$
MTL-CNN	$0.68\% \pm 0.06\%$	$3.12\% \pm 0.06\%$	$3.58\% \pm 0.14\%$

The result shows that across all the tasks, the MTL-CNN model outperforms other methods. Additionally, the value of the standard deviation on all trials is also small indicating the robustness of the model. The method using geometrical descriptors even outperforms the two-point statistics especially for the predicting of rubbery modulus and $\tan\delta$ peak, which suggests that those hand-crafted descriptors designed by material express are really good at capture the characteristics of the microstructure and informative to the bulk properties of the composites. Compared with method using geometrical descriptors, deep learning model improves the accuracy for prediction of glass modulus by as much as 45.2%, rubbery modulus by 34.2%, and $\tan\delta$ peak by 19.7%. In addition, to demonstrate the data volume is sufficient, the model is trained using different portion of data and the accuracy of the model is evaluated. Figure 31 shows the accuracy of the model (in MAPE) as a function of different training data portion. Based on the plot, for three property predictions, there is significant accuracy improvement when the training data increases from 0.2 to 0.6 then later the curve gradually flat out, which indicates the data size is sufficient to describe the variations across the whole space.

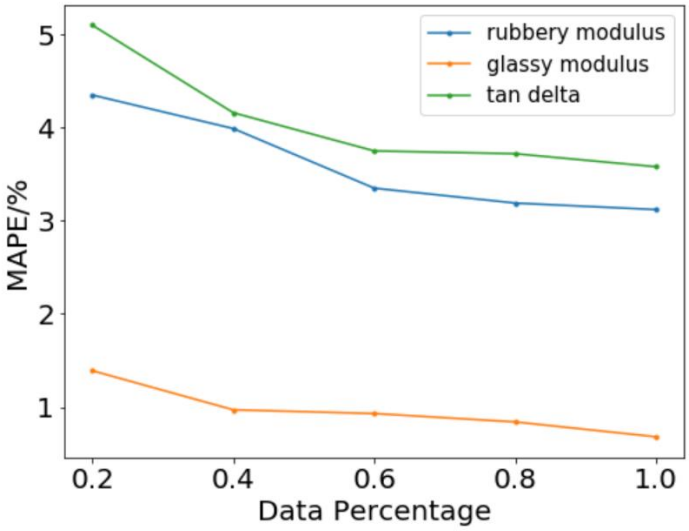

Figure 31. Plot shows the accuracy of the prediction for three property values as a function of different training data size (in percentage).

4.2.8 Model interpretation

As given in Table 4, deep learning model outperforms other traditional machine learning model for all the three tasks. However, deep learning model is relatively hard to interpret compared with traditional method using hand crafted features. In another word, for traditional machine learning method, it is easy to extract what is the impact of every feature to the target while deep learning model works as a black box with complex architectures with millions of parameters. Although it is not possible to interpret every parameter in the deep learning model, it is worth the effort to understand what the model learns and whether it captures some physical knowledge behind. Therefore, in this section, by modifying the inputs to the deep learning model, we try to interpret what the model learns from physical point of view.

In section 2, it has been shown that the glassy modulus increases monotonically as a function of volume fraction. In another word, for a specific microstructure, more fillers will lead to a larger glassy modulus value. This also suggests that the filler phase have larger impact on the glassy modulus compared with that of the matrix. We would like to validate whether the proposed model captures this physical knowledge. To do so, the original microstructure input is modified by randomly removing a number of pixels from filler or matrix respectively and the modified response could be compared. The removing process is achieved by setting the value of that pixel to 0, which eliminates the activation of the model from that pixel and therefore the contribution to the target value. Figure 32 shows an original image and the modified images after removal different number of pixels from filler or matrix. To compare the impact of filler phase with the matrix phase, two sets of modified images are generated, one is to randomly remove pixels from filler and the other is to randomly remove pixels from matrix. Then later, the modified responses from those two sets are compared with the original response. The change of modified response is evaluated according to the residue equation below:

$$r = \frac{|\bar{y}-y|}{y} \times 100\% \tag{15}$$

where $\bar{y}$ is the predicted value from the deep learning model and y is the true value from FE model.

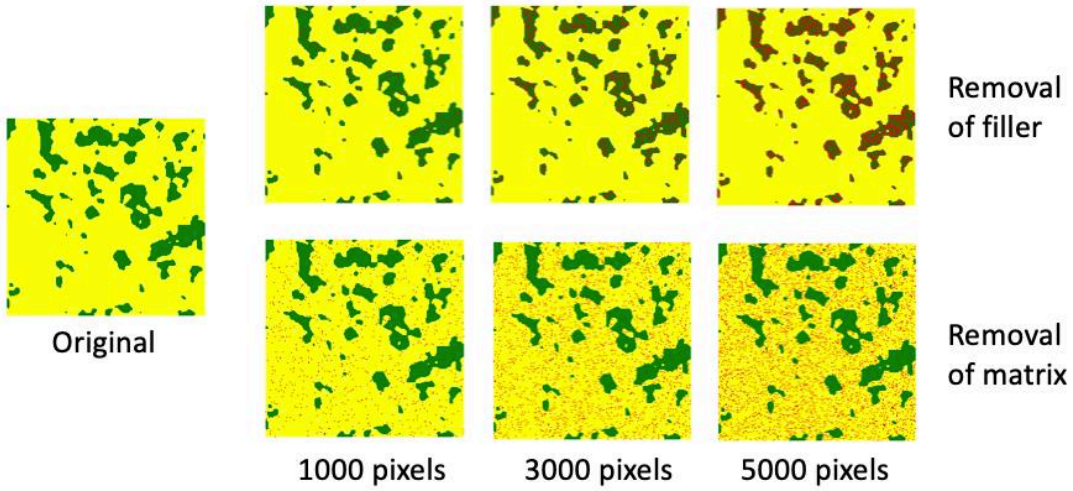

Figure 32. Visualization of the original microstructure and the modified microstructure with different number of pixels removed from the original image. Blue represents the matrix region while green represents the filler region. The pixels with red color are removed by setting value to zero.

Figure 33 shows the contribution of matrix and filler in the trained deep learning model by plotting different removed pixel number with the residua. The plot is based on a single testing sample and the experiments is conducted ten times for accuracy and stability. The plot shows that as the increase of removal pixels for either from matrix or filler phases, the residual value increases monotonically, indicating that the deep learning model has lower accuracy if more pixels are eliminated. More importantly, the residual value from the filler is much higher compared with the residual value from the matrix, which suggests that the contribution from the filler is higher than matrix in determination of glassy modulus. This finding, corresponding to the knowledge illustrated earlier, were not explicitly introduced to the deep learning model during the training process. The proposed deep learning model is able to learn and capture this physical knowledge during the training process. Additionally, in order to show that this finding holds for other samples,

the same experiments is done for ten random chosen testing samples and the average residual is calculated across those samples. Figure 34 shows the residual for ten different samples. The plot shows that on average, the effect from the filler is higher than the matrix in determination of glassy modulus. Therefore, the trained deep learning model is able to implicitly learn and capture these physical knowledges for accurate predictions.

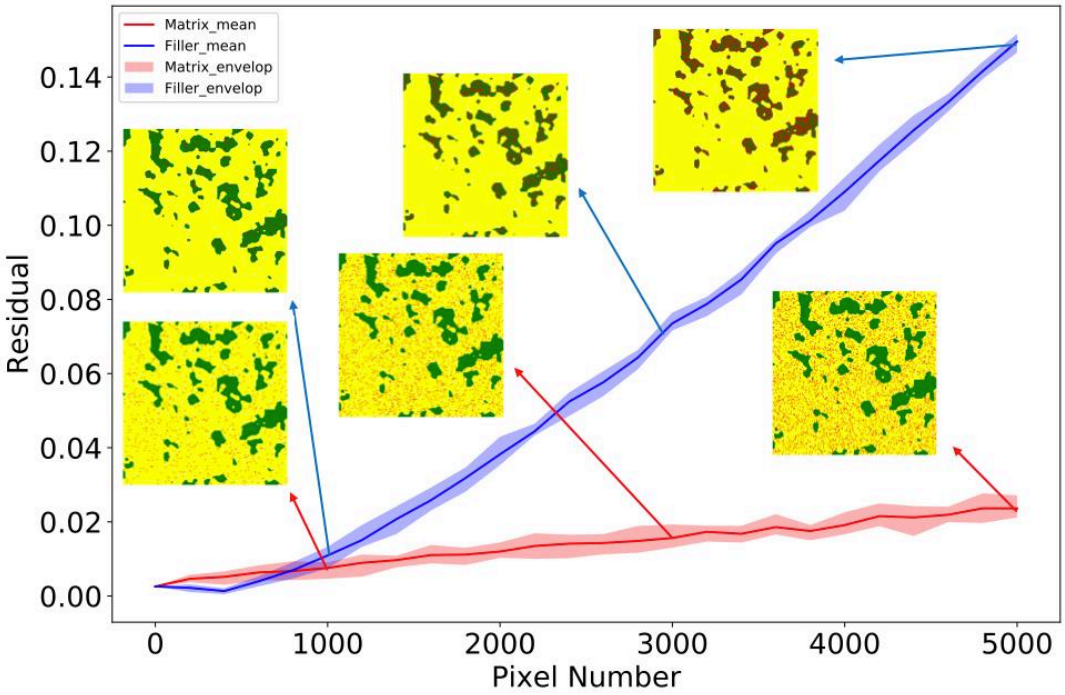

Figure 33. The plot shows the residual as a function of removed pixels for different material phases on a single testing sample. The experiments are conducted ten times for robustness and accuracy. The colored area shows the distribution of values for ten trials and the solid line shows the average.

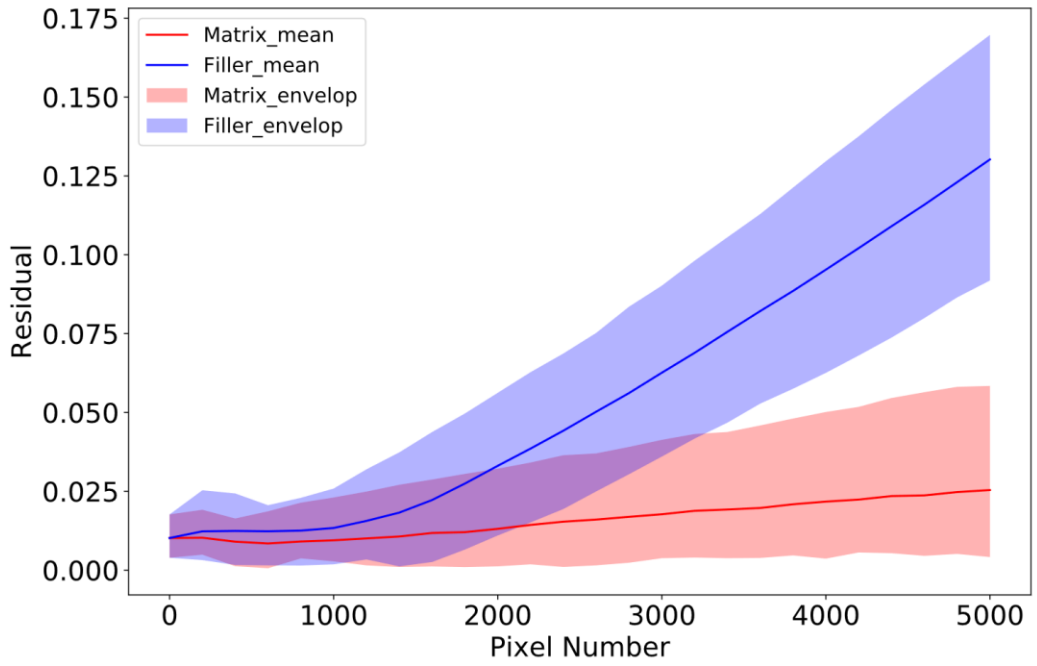

Figure 34. Average residual plots for ten different testing samples. The colored area shows the distribution of average residual for different samples while the solid line shows the mean value.

4.3. Conclusion

In this paper, a multi-task convolutional network is proposed to predict mechanical properties of polymer nanocomposites using microstructure images. A computational data set with size 11000 is generated using finite element simulations. The performance of the model is benchmarked against two other state-of-the-art approaches using two-point statistics and structural geometrical descriptors. The result shows that the proposed deep learning model improves the accuracy for prediction of glass modulus by as much as 45.2%, rubbery modulus by 34.2%, and tan δ peak by 19.7%. Additionally, by modifying the inputs to the deep learning model, we have shown that the deep learning model is able to capture physical understandings through learning,

which are not explicitly introduced to the model before-head. The proposed deep learning approach is a feature engineering free, high accuracy and generalization and interpretable model to study the structure-property linkage in polymer nanocomposites.

Because of the feature engineering free characteristic, this method can be applied as a generalized model to build structure-property linkage for other material systems and properties, e.g. meta materials. Although this work is primarily relied on computational data, similar study can be accomplished in the future if more experimental data are available. Future study can also work on training deep learning model using multi-phase images including interphase layers with gradient properties. Then the trained model can be applied to extract other material insights including the impact of gradient interphase layers and help to better understand the role of interphase in polymer nanocomposites.

Chapter 5. Nanocomposite synthesis information extraction from scientific literature using Natural Language Processing (NLP)

5.1. Introduction [2]

Material Genome Initiative (MGI) has dramatically drive the developments of open-source web-based material data platforms, which expediates the material development and design. For example, in the metals community, several mature materials databases are widely used, such as the CALPHAD database for alloy phase diagrams which is over three decades old [14]. Other efforts, such as the Materials Project [15], leverages high performance computing (HPC) to accelerate materials discovery, particularly in the field of next generation batteries using simulated datasets of inorganic compounds [13]. OQMD has enabled researchers to quickly obtain density functional theory (DFT) calculated thermodynamic properties and apply tools to visualize phase diagrams and crystal structures for many complex alloys and compounds [16, 17]. In the polymer nanocomposite filed, although the development of data-driven platform is not as well-developed as in the metal field, there exists several examples of online databases that contains sufficient data sets and easy-accessible web interface. These include the CRC POLYMERSnetBASE by Taylor and Francis Group [19], the Polymer Property Predictor and Database (PPPDB) by University of Chicago and the PolyInfo database from NIMS of Japan [18]. All these data resources distribute curated polymer data from publications and polymer handbooks, with detailed annotations of

[2] This section is from published paper on *APL Materials* 6.11 (2018): 111108.

chemical properties and characteristics. However, there are limitations for those data resources: for PolyInfo, the lack of API access prevents the application of user-defined search and exploring of data; the CRC POLYMERSnetBASE requires paid access; and PPPDB covers only a few properties (notably the chi parameter and glass transition temperature) and is focused on polymer blends. Specially, as an exclusive data resource for the nanocomposite research community, we developed NanoMine as a prototype framework of an online, open resource system for nanocomposite materials data sharing, analysis and material design [46]. It includes a nanocomposite database with online access and REST API, a suite of web-based tools for statistical analysis of microstructure images and material properties, as well as physics-based modeling and simulation of nanocomposite properties. Currently, the NanoMine database contains 182 papers, over 1000 unique samples, covering over 100 types of polymers/particles with over 20000 data points and is continuously expanding.

For most of the material data resources, including NanoMine, the data curation process is still relied on a human extraction procedure, which requires a curator with domain knowledge to read through the paper and extract all the information based on pre-defined schema. This manual curation process is expensive, error-prone and labor-intensive. Based on our experience, for data curation in NanoMine, it requires a full-day work to extract all the sample information from one paper and input those into the database. Plus, with the continued growth of new publications, it is becoming increasingly difficult to maintain and update manually curated databases.

Natural language processing (NLP)，which enables automatic extraction of data from literatures using statistics and computation techniques，provides a possible alternative for manual

curations. NLP is a subfield of machine learning and computer science focusing on using computational techniques to learn, understand and produce human languages. The applications of NLP include machine translation, information extraction, social media mining and so on [112-114]. Traditional NLP approaches primary rely on rule-based method by hand-crafting a set of rules. And the quality of the NLP models is heavily dependent on the quality of designed rules. However, due to the high variability and ambiguity of human language, writing rules of human languages is non-trivial and difficult. Starting from the late 1980s, NLP gradually transformed to apply machine learning algorithms to understand the grammar [115, 116].

In chemistry and biomedical domain, there are many well-established NLP tools. ChemDataExtractor [44], is one of the tools that can extract chemical information from the scientific literature. Combining different NLP techniques, the system provides a complete pipeline for the automated extraction of chemical entities and their associated properties, experimental measurements etc. Kim el al. [41, 42] developed a framework to automatically compile material synthesis information from literatures using different NLP techniques for metal oxide systems. While there are many well-established NLP tools in other domains, there is no such tool available in polymer nanocomposite filed probably because of the complexity and high dimensionality of nanocomposite data space, accounting for all detailed information in chemistry, processing structure and property spaces.

However, in the polymer nanocomposite domain, such NLP model has not been developed because of the complexity of nanocomposite processing, characterization and measurements. By applying recent machine learning and natural language processing methods, an end-to-end

framework to extract material processing and synthesis information from full-length journal publications is proposed, developed and evaluated.

5.2. Methodology

5.2.1 Nanocomposite Schema[3]

To appropriately capture the full suite of possible data for nanocomposites, there are six major sections in the NanoMine schema. In each category, many hierarchically related fields have been defined to store sufficient detail on nanocomposite data to enable subsequent data analytics and design tools. The structure and fields have been designed using a test set of published papers on nanocomposites and have been further refined with additional curation of data into the data resource. The top-level categories in NanoMine and the description of the subfields organized underneath them are as follows:

- **Data provenance**: Metadata of the source of the literature guided by Dublin core standards[104]. The essential metadata includes the DOI of the cited source, author, title, keyword, time and source of publication, etc. This section of the schema supports both published and unpublished datasets. For published datasets, we have developed a robust automatic DOI information retrieving tool [117]that fills out the data source information for users and enhances the data quality by eliminating entry errors. NanoMine currently stores data from 12 publishers and the tool supports all of them.

[3] This section is from published paper on *APL Materials* 6.11 (2018): 111108.

- **Materials Composition**: Characteristics of the constituent materials, including the polymer matrix, the filler particle, and surface treatments for polymer and/or particle constituents. Bulk matrix and filler properties can be entered, along with compositions (volume/weight fraction) and pre-treatment (for example, grafting and other surface treatment methods).
- **Processing**: Extracted sequential description of chemical synthesis and experimental procedures. Currently, three major sub-categories are included: Solution Processing, Melt Mixing, and in-situ polymerization. Detailed information at each processing step, such as temperature, pressure, and time, can be entered.
- **Characterization**: specification of material characterization equipment and methods and conditions used. This information includes details on common microscopic imaging (SEM, TEM), thermal and electrical measurement, mechanical property measurements, as well as nano-scale spectroscopy.
- **Properties**: measured data of materials performance and response. Properties include mechanical, viscoelastic, electrical, thermal and volumetric properties. Format of data can be scalar or higher dimensional, such as in 2D plots or 3D maps.
- **Microstructure**: contains nanophase spatial dispersion and topological information from micro and nano-scale imaging. Multiple images can be archived to document microstructure in nanocomposites. Unlike traditional journal articles, where typically one "representative" image is recorded, we provide the ability to archive sets of images, which will increase robustness of statistical descriptors of microstructure [64, 98] and the correlations [118] that can be gleaned between structure and processing or properties.

Geometric descriptors can also be entered into this section to describe the statistical distribution of microstructure.

The full representation of the current schema in the XML tree can be found on GitHub [119] and on the NanoMine website [120]. We utilize the NIST Materials Data Curation System (MDCS) platform [121] as the backbone of the NanoMine database. Experimental nanocomposite samples have been effectively populated into the NanoMine database using this schema. Figure 35 shows an example of a populated XML document of a graphite-PMMA nanocomposite sample. The hierarchical structure indicates some basic relationships across the parameters. For example, the “Particle Size” can be defined specifically as a child-element under “Filler”, with value and unit as its subsequent child elements. We will see later that mapping this schema to an ontology will provide an even more flexible framework to handle both simple and complex relationships far beyond the parent-child concept.

Figure 35. Populated XML tree for a given sample in NanoMine. As highlighted in two red boxes, "Particle Size", with two sub-elements "value" and "unit", is a child element of "Filler", which is in the "Materials Composition" upper level category.

The non-relational format of the schema enables efficient editing and versioning. The schema is a living system and will continue to evolve with curation. For example, solid state processing is a processing method that is not currently included and will be added to an upcoming version of the schema as data from papers utilizing it are curated. Editing the schema can be conveniently performed by retaining the original hierarchical relation and appending any new elements as child or parent to corresponding fields. For example, a new processing condition type (eg, solid state processing) can be inserted along with associated new physical quantities (eg, residence time) into the existing schema, retaining existing representation of all other previous terms to ensure backward compatibility.

Figure 36. Parallel coordinate plot of Nanomine samples for selected parameters (polymer, filler, loading, Tg and characterization method)

Figure 36 illustrates a parallel coordinate plot showing an overview of Nanomine samples for some selected dimensions/parameters. The color of each line indicates the number of distinct properties recorded for each sample. From the plot, it can be seen that NanoMine has a wide range of interconnected data on each nanocomposite sample by using our designed NanoMine Schema. The rich information set will enable robust discovery and development of search and visualization tools.

5.2.2 NLP framework

Although a fully automated data injection is not feasible at the current stage given the complexity of languages and writings, a semi-automated curation procedure assisted by NLP is developed to ease the manual labor and improve the accuracy. Based on our experience on manual

curation, it is found the extraction of material processing information most time-consuming because that information is usually embedded in pure text and requires an experienced curator with domain knowledge to given a thorough reading in order extract the data. On the other hand, the property or microstructure information are relatively easy to obtain as those are usually presented in the form of a table or plot. Therefore, as the first step toward automated curation via NLP, a toolkit to extract material processing information from literatures is valuable and could assist the manual curation process.

To do so, a pipeline of machine learning models is proposed working at different sections and levels of papers (e.g. abstract, paragraphs, sentences) to gradually extract the relevant processing information from full-length papers. The proposed NLP framework in shown in Figure 37**.** The paper collection can be done through using Application Programming Interface (API) from different publishers, which enables retrieving articles in either PDF or plain text format. In this way, a corpus of journal publications can be obtained. However, not all the collected papers contain the nanocomposite data we desired so the NLP model 1 is built to filter out the irrelevant papers. Secondly, NLP model 2 serves as a paragraph classifier selecting the relevant paragraphs containing the material processing information from the whole paper. Thirdly, the processing goes down one more level and focuses on understanding the meaning of each sentences. Inspired by the definition of schema, individual sentences are further classified into four different categories (e.g., material characteristics, experimental action, irrelevant information etc.) based on their semantic meanings. Lastly, taking advantages of outputs from grammar parser and combining with different heuristic rules, the exact experimental procedure and relevant conditions can be extracted.

Due to the lack of API accessibility for paper downloading, in the following section, details about NLP model 3 and NLP model 4 as well as the results are presented while the building of the first two NLP models requires future efforts in order to complete the whole framework.

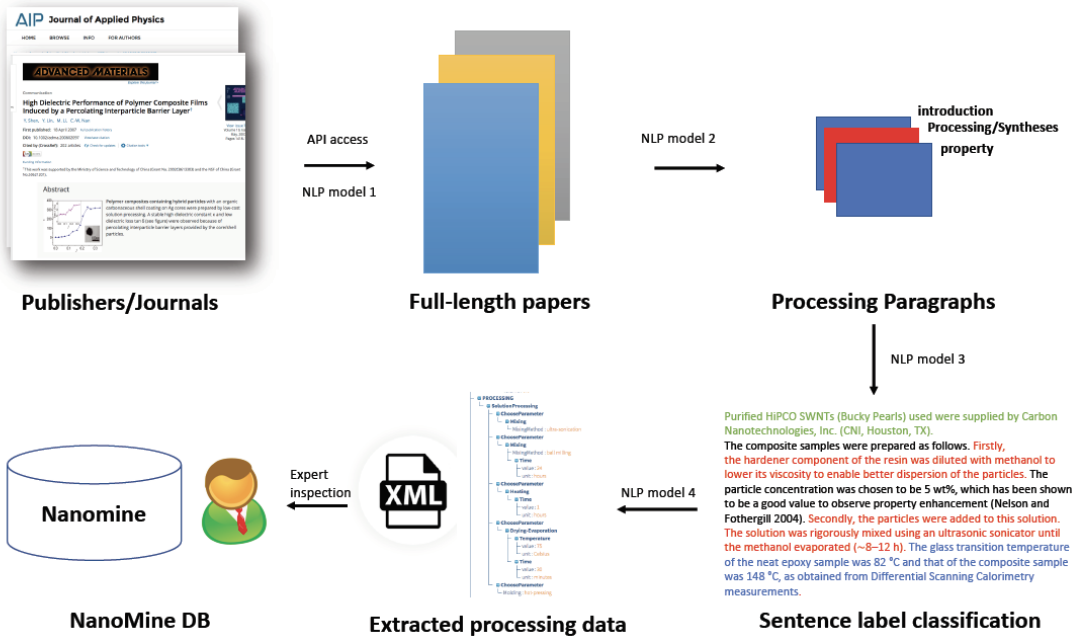

Figure 37. Workflow of applying NLP to extract material synthesis information from literature. The first NLP model is built to filter out the irrelevant papers; NLP model 2 is constructed working as a paragraph classifier selecting the material processing paragraphs. Then the processing goes down one more level and focuses on understanding the meaning of each sentences by classifying each sentence into different categories. Lastly, taking advantages of outputs from grammar parser and combining with different heuristic rules, the exact experimental procedure and relevant conditions can be extracted.

5.2.3 Word embedding

The workflow to solve most of the NLP tasks is similar to solve machine learning problems, which typically includes feature extraction, model selection and evaluation as shown in Figure 38. NLP is also a machine learning problem other than the data is words and letters instead of numbers. Given the fact that the machine learning models such as logistic regression or decision trees usually takes the input in numerical forms, the first step of NLP is to map the words and phrases into real

numbers, which is also called word embedding. There are two types of methods for feature extraction from text: count-based methods, such as bag-of-words and term frequency-inverse document frequency (TFIDF) [114, 122], and continuous vector representations, such as word2vec and Glove [123, 124]. For count based methods, the bag-of-words model is a simplified representation using the frequency of occurrence of each word [122]. For example, given two simple text documents: (1) John likes cat. Mary likes cat too. (2) John also likes dog. If we count the occurrence of each word, we would get (1) [John: 1, Mary: 1, likes: 2, cat: 2, too: 1]; (2) [John:1, also:1, likes:1, dog:1]. Then we could form a word library by union all the unique words from both documents: library = [John, Mary, likes, cat, too, also, dog]. Lastly, for every word in the library, if we use the term frequency, we will obtain the bag-of-words representation of each document (1) [1,1,2,2,1,0,0] (2) [1,0,1,0,0,1,1], which can be considered as features and applied for machine learning models. However, this simplified representation may not be the best representations. Common words such as “I”, “you”, “the”, “a” are always words with highest terms, however, these words do not carry useful information. Therefore, TFIDF is proposed to solve this problem by normalizing the term frequencies by inversing the document frequency [125].

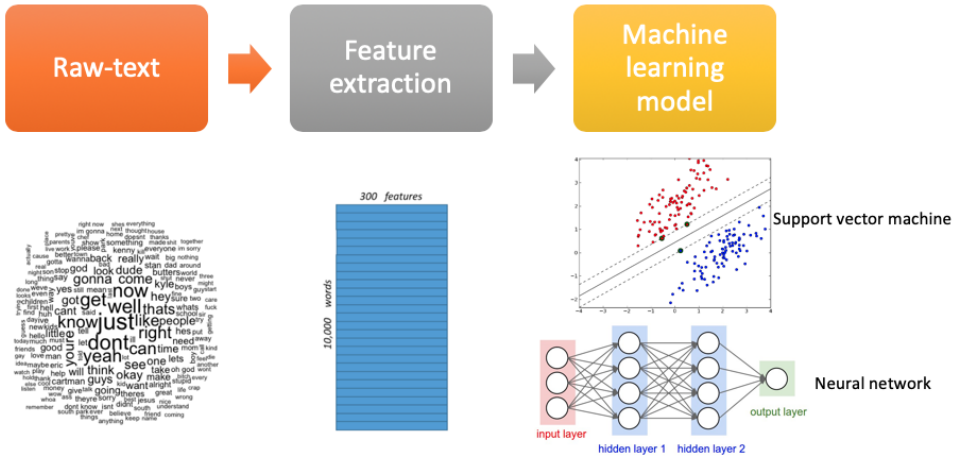

Figure 38. A common workflow of NLP. The method starts from extracting features from raw text, followed by applying machine learning on top of the extracted features.

Even though count-based word representations are simple and robust if trained on a large amount of data, this representations loss essential information in languages: the notion of semantic meaning of words, the similarity of words and the connections. Additionally, in most of the NLP tasks, the amount of in-domain data is limited. Instead, researchers developed more advanced techniques using continuous vector representations. One of the popular vector representations is word2vec, which is a continuous representation that is learnt by a two-layer neural networks. This vector representations can be learnt in two ways: either to use context to predict target word (also called continuous bag of words or CBOW) or to use word to predict context (also called skip-gram) [123]. Compared with other methods, word2vec has greatly improved the quality of the learned word representations, which is able to capture the semantic meaning of each word and learn implicitly the similarity between different words. Although the continuous vector representations are better compared with count-based methods, the continuous vector representation need to be obtained by training on a relatively large corpus, which is not available for our study because of lacking API access to obtain sufficient amount of training papers. Instead, we take advantages of

a transfer learning strategy where we directly apply pre-trained embeddings coming from other applications. There exist several different pre-trained models published by different institutions that are trained on millions of words from Wikipedia, twitter, or newspapers. For example, Pennington et al. [124] from Stanford published GloVe word vectors [124]. Google also published pretrained embeddings using word2vec . However, those pretrained embeddings cannot be directly applied to material research as the corpus are different. Material research contains a lot of domain specific words and terminologies that may not be covered in daily documents. A very recent research bridges this gap by training word embedding using material research articles. Kim et al. [41, 42] trained word embedding on 640000 full-text metallic material synthesis articles and obtained vector representations. Although the material system is different, our scenario is still very similar to theirs (both investigating experimental procedures for material synthesis) and those word vectors from metallic material synthesis is transferable to our application.

In order to examine the quality of the pre-trained vectors, we try to identify 'similar' words by calculating the cosine similarity and ranking according to that score. Table 5 shows the most similar words according the cosine similarity score when the target word is "silica". These words have similar semantic meanings or functionalities compared with the target word, which demonstrates that the pre-trained word vectors can implicitly model the connections between different words.

*Table 5. Top 5 most similar word with '**silica**' according to cosine similarity score.*

word	score
Silica	0.596

brucite	0.513
boehmite	0.509
silicic	0.503
titania	0.502

5.2.4 Sentence Classifier using attention networks

After encoding each word into vectors, those features are ready to be applied to a machine learning model for classification. Inspired by NanoMine schema [46, 126], each sentence from the processing paragraph is classified into one of the four categories: material constituents, material properties, experimental action and not relevant. More details about each category is shown in Table 6**.**

Table 6. sentence label, interpretation and examples.

label	interpretation	examples
material constituents	Filler/polymer chemical name and manufacturer	Aniline and ammonium peroxydisulfate (APS) are purchased from E Merck (India). Tritherm® B 981-N-42, a (PAI) resin, was chosen as the matrix material.
material properties	Filler/polymer properties such as density, glass transition temperature.	The COC was Topas® 8007S04 (Ticona GmbH, Germany) with a density of 1.02g/cm3 and a melting temperature of 190–250◦C.

experimental action	the experimental procedures to synthesis the nanocomposites	Then, the appropriate amount of hardener is poured into the beaker, mixed vigorously with hand for few minutes and poured into the mold.
not relevant	All the other sentences that does not contain useful information	The surface-initiated RAFT polymerization process has been previously documented and is briefly discussed here.

The process to classify each sentence into the above four categories is a text classification problem. Traditional approach on text classification relies on combing count-based embedding methods with linear model or kernel methods such as logistic regression or support vector machines (SVM). Then the NLP gradually transforms to deep learning, such as using convolutional neural networks and recurrent neural networks. Yang et al. [127] presented a novel hierarchical attention networks (HAN) for text classification that better represents the structure of the documents. The HAN is designed according to two insights from document structure. The first one is that the documents have a hierarchical structure (words form sentences, sentences form a document). The HAN is constructed to first generate sentence representations from words and then to aggregate sentence representations to form a document representation. The second insight is that not all the words and sentences are equally informative.

Inspired by the HAN, Figure 39 shows the proposed attention network for sentence classification [127, 128]. Taking the pre-trained embeddings, the word representation is feed to a bidirectional gated recurrent unit (GRU) to obtain the encoding of each words. Then an attention mechanism is applied to pay more attention on the words that carries significant meanings. Finally,

the averaged vectors according to all weighted values are feed to the final softmax layer for prediction.

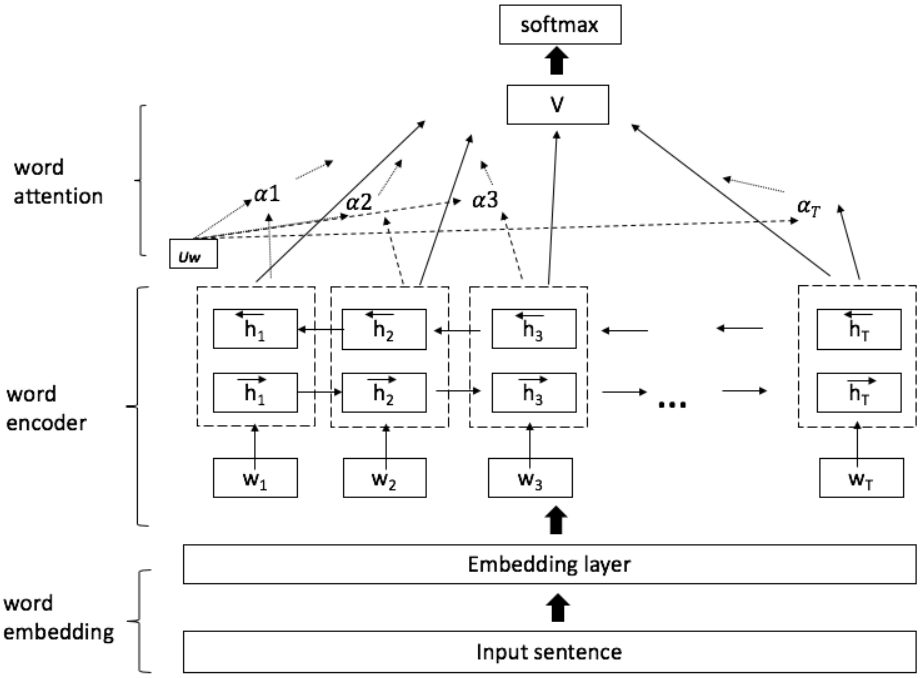

Figure 39. Attention network for sentence classification

5.3. Result

The sentence classification model is trained on a dataset with 2000 sentences that is manually labeled by material experts. Each sentence is assigned with a unique label based on its semantic meanings. The performance of the model is benchmarked on other approaches using linear classification model such as logistic regression (LR) and support vector machine (SVM). The evaluation metric is F1-score. The results are given in Table 7. Result comparison for methods using different machine learning models and embeddings. The result shows the F1-score.**.** The results compare the performance of classifier using different machine leaning models and embedding methods. The results show that the linear model works better together with count-based

embeddings such as bag of words (BOW) or TFIDF, while the deep learning model shows better performance using learnt continuous representation such as word2vec. Overall, the performance of the deep learning model is the best among other benchmark methods, however, the performance of the deep learning does not exceed a large extent compared with traditional method. This is probably because for this study, the size of the training data is small. An improved performance of the deep learning model should be expected if more training data are available.

Table 7. Result comparison for methods using different machine learning models and embeddings. The result shows the F1-score.

Model	F1-score
LR + BOW	0.81
LR + TFIDF	0.79
LR + word2vec	0.80
SVM + BOW	0.82
SVM + word2vec	0.83
SVM + TF-IDF	0.85
Attention NN	0.84
Attention NN + word2vec	0.88

Next, we would like to demonstrate the interpretation of the model by plotting the weights from the attention layer and show that the model is able to catch the critical words and assign them more weights. The visualization is shown in Figure 40. This sentence is from the testing set that has not been exposed to the model during training. The figure shows the words together with the importance score learnt by the model. From the result, the model catches the critical words such as ‘were’,’mixed’ and assign them much larger weight values compared with other more common words such as ‘at’,’the’ etc. The ability to recognize important words in a sentence is very like to

how human interpret the meaning of sentence, which is the reason why the model is called an 'attention' model.

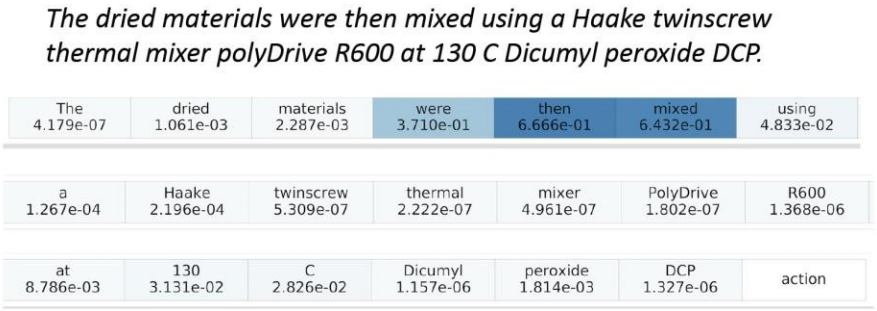

Figure 40. Visualization the weights of the attention layer. A darker color represents a larger weight and more importance of the word.

After the sentence classifier, different parsers are built to extract the detailed experimental procedures and parameters. The method is primary based on combing syntax analysis with rule-based grammars. The workflow for the parser is given in Figure 41.

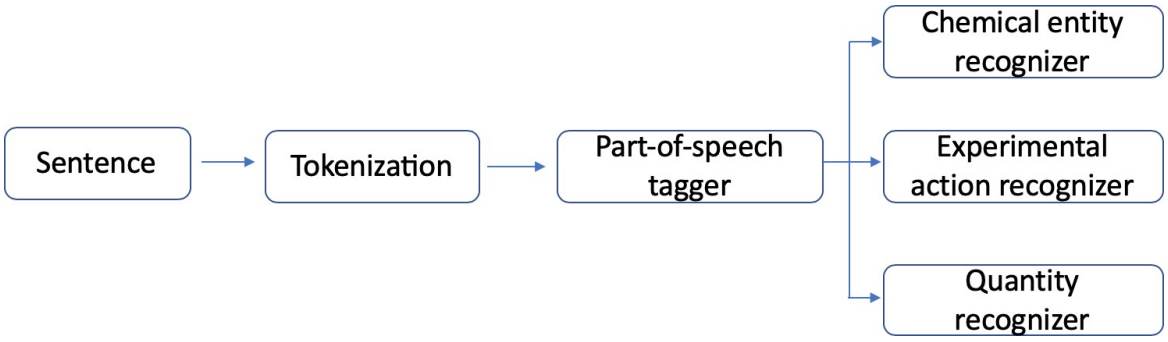

Figure 41. Workflow of the parser to extract detailed material processing procedure and parameters

Taking an example sentence, *'The polycarbonate suspension was stirred for 0.5 h at 55 °C under the protection of N2, after cooled to room temperature'*, we demonstrate how the parser works step by step.

Tokenization: Tokenization is the process split the sentences into individual words and tokens. After the tokenization, the output is *[u'The', u'polycarbonate', u'suspension', u'was', u'stirred', u'for', u'0.5', u'h', u'at', u'55', u'\xb0C', u'under', u'the', u'protection', u'of', u'N2', u',', u'after', u'cooled', u'to', u'room', u'temperature', u'.']*

Part-of-speech (POS) tagger: POS tagging is the process to grammatically tag each words or to generate word-category disambiguation; simplified tagging in the identification of words as nouns, verbs, adjectives, adverbs, etc. The output from the POS tagger is *[(u'The', u'DET'), (u'polycarbonate', 'NOUN'), (u'suspension', u'NOUN'), (u'was', u'VERB'), (u'stirred', u'VERB'), (u'for', u'ADP'), (u'0.5', u'NUM'), (u'h', u'NOUN'), (u'at', u'ADP'), (u'55', u'NUM'), (u'\xb0C', u'NOUN'), (u'under', u'ADP'), (u'the', u'DET'), (u'protection', u'NOUN'), (u'of', u'ADP'), (u'N2', 'NOUN'), (u',', u'.'), (u'after', u'ADP'), (u'cooled', u'VERB'), (u'to', u'PRT'), (u'room', u'NOUN'), (u'temperature', u'NOUN'), (u'.', u'.')]*.

Phrase chunking: this is the process to identify phrases by combining tokens. In our case, we focus on identifying noun phrases. The chunking mainly relies on rule-based grammar. An example grammar rule to capture noun phrases can be: [<DET>?<ADJ>*<NOUN>*],which means the noun phrase has a pattern combining articles, adjectives and noun. By carefully designing these grammar rules, noun phrases such as 'The polycarbonate suspension', 'room temperature' is captured.

Chemical entity recognizer: To capture the chemical entities, two approaches are applied. A chemical corpus: CHEDNER (10000 PubMed abstracts with manual annotated chemical entities is taken and worked as the first step to pick up the chemical entities [129]. For other chemical entities such as polymer names, rule-based grammar is designed. Utilizing the proposed method, chemical entities can be extracted. After the processing of chemical entity recognizer, 'The polycarbonate suspension' and 'N2' can be identified as chemical entities.

Experimental action recognizer: Similar to the idea to pick up chemical entities, rule-based grammar is designed to pick up the exact processing steps such as mix, stir, sonicate etc. The grammar rule for identifying experimental actions is "*.*extrud.**|.*sonica.**|.*stir.**|.*mill.**|.*centrifu.**|.*dissolv.**|.*mix.**|.*deaer.**|.*cool.**|.*heat.**", which is the combinations of root words of different experimental actions. After this step, 'stirred' and 'cooled ' are identified as experimental actions.

Quantity recognizer: the objective of quantity recognizer is to pick out the quantity parameters such as time, weight and temperature. To do this. Rule-based grammar is designed. The grammar to extract the time phrase is *"[0-9]+.*h|[0-9]+.*min",* which suggests a pattern including numerical character at the beginning followed by time units such as hour or minutes.

After the processing of the parser, words and phrases with detailed experimental actions and chemical entities can be tagged using different colors as shown in Figure 42**.** This tagging provides support and makes it easier for manual curators to input data to the database.

'The polycarbonate suspension was stirred for 0.5 h at 55 °C under the protection of N2, after cooled to room temperature.'

Figure 42. Example output from the parser with different colors marking different types of information such as experimental actions, chemical entities etc.

5.4. Conclusion

By applying recent machine learning and natural language processing methods, an end-to-end framework to extract material processing and synthesis information from full-length journal publications is proposed, developed and evaluated. The framework consists a series of NLP models that process information from different paper sections. A paragraph classifier is first built to select the relevant paragraphs which contains the processing information from the whole paper. Sentences from the processing paragraphs are then further classified into four different categories using a hierarchical attention neural network based on their semantic meanings. The network is trained against a set of over 100 human annotated literatures, labelled by material scientists while reading through the paper. Lastly, different heuristic rules are applied according to the sentence categories to extract the exact material processing actors and parameters.

Although at the current stage, the training corpus is still small, however, taking advantages of manual labelled sentences from 100 papers, the proposed models and methods is still proved to be a good candidate to assist manual-based curation and improve the efficiency of data ingestion for the database.

Chapter 6. Conclusion and future works

6.1. Summary of contributions

NanoMine, a polymer nanocomposite resource consisting of a curated database, search and visualization tools, data analysis tools, and simulation tools was built by an interdisciplinary research team in order to accelerate materials design. To further expand the functionality of NanoMine for adoption by the larger nanocomposite research and design community, efforts have been undertaken to enhance three aspects of the data resource. The first thrust is to build more analysis and module tools that better model material behavior and exapand the toolbox for material design. The second perspective is to carry out materials science case studies using NanoMine data aligned with the materials science paradigm of elucidating processing-structure-property relationships in order to better understand the physics underlying material behavior. Last but not the least, it is essential to put continuous efforts on enhancing data curation, which expands the size of database and enables development of more systematic studies using data science techniques. Corresponding to these three thrusts for advancing the capabilities of NanoMine, this dissertation highlighted three research tasks: 1) building more robust and accurate computation models, 2) mining structure-property relationships using NanoMine, and 3) designing an NLP framework to assist manual data curation. To the best knowledge of the author, the work of inverse determination of interphase properties using adaptive optimization in Chapter 3 was the first work that combines FEA models with optimization methods to automatically search the optimal interphase properties for polymer nanocomposites. In addition, the proposed deep learning model for structure-property modeling presented in Chapter 4 was the first work that builds quantitative relationships via

feature-engineering free, interpretable, and multi-tasked deep learning model in the context of polymer nanocomposite materials. Additionally, the proposed NLP toolkit for material synthesis extraction described in Chapter 5 was the first implementation of NLP in the field of polymer nanocomposites. More detailed contribution about each task is given below:

Task 1. An inverse approach identifying interphase properties in polymer nanocomposites:

This work demonstrated a consistent and efficient approach for identifying the interphase properties in polymer nanocomposites by solving the inverse problem using adaptive optimization, improving the prior work which was limited in accuracy and efficiency by the laborious manual iteration process. Using the proposed automated approach and given experimental bulk property and microstructure data, the interphase properties in dielectric and viscoelastic studies were determined automatically. The results demonstrate that only tens of iterations are required for the method to identify the optimal shifting factors and interphase properties to achieve a good fit with the experimental data, shedding insight into the profile of interphase properties near a confining interface. Comparisons among different existing searching algorithms showed the ability of our method to reduce the computational cost and improve searching efficiency.

This method can be used as a generalized automated approach to determine the interphase properties in polymer nanocomposites. The framework is also flexible and can be applied to other computational models to inversely determine the required interphase properties. Additionally, the method is very efficient compared with the manual-fitting process, which will facilitate further investigation by data-driven analysis where many hundreds of samples may be required.

Task 2. mining structure-property relationship using NanoMine:

This work first demonstrated the application of NanoMine on finding trends and facts that is already known to material scientists, which further motivated the subsequent systematic investigation of structure-property relationships using computational data. The first part of the analysis focused on qualitative relationships between microstructure descriptors and mechanical properties, resulting in new findings about the interplay of interphase, volume fraction, and dispersion. A multi-tasking convolutional network was proposed to predict mechanical properties of polymer nanocomposites using representative microstructure images. The performance of the model was benchmarked against two other state-of-the-art approaches using two-point statistics and structural geometrical descriptors. The results indicated that the proposed deep learning model improves the accuracy for prediction of polymer modulus by as much as 45.2% and 34.2% in the glassy and rubbery regimes, respectively, and for $\tan \delta$ peak prediction by 19.7%. Additionally, by modifying the inputs to the deep learning model, it was shown that the deep learning model is able to capture physically meaningful understanding through learning, without explicit introduction to the model beforehand. The proposed deep learning approach is a feature-engineering-free, high accuracy, generalizable, and interpretable model to study structure-property relationships in polymer nanocomposites.

This work demonstrated the ability of NanoMine in testing hypotheses and providing insights for understanding material mechanisms. This analysis framework can be extended to test other hypotheses put forward by materials scientists and to help elucidate unknown mechanisms behind material behavior. Additionally, due to the feature-engineering-free, interpretable, and

robust characteristics of the proposed deep learning model, it can be transferred to other material system classes, e.g., metamaterials.

Task 3. Extracting material synthesis information from scientific literatures using NLP:

In this work, by applying advanced natural language processing and machine learning models, a framework consisting various machine learning models was proposed to extract material processing and synthesis information from scientific literatures. A series of NLP models were trained and applied to process information from different paper sections. Firstly, a paragraph classifier was trained to select material processing paragraphs from the full-length paper. Secondly, each sentence from processing paragraphs was further classified into different categories using a deep learning model based on their semantic meaning. Lastly, rule-based parsers were built to extract the exact experimental actions and parameters. The method was trained and evaluated on a set of over 100 human annotated papers and shown good performance. Although it is acknowledged that the current training corpus is relatively small, the proposed framework and models were still proved to be a powerful tool to assist manual-based curation and improve the efficiency and accuracy of data ingestion for the database.

6.2. Future work

Interphase property prediction based on material characteristics and processing information

In this dissertation, an inverse method to identify interphase properties has been developed through combining FEA with Bayesian optimization. However, this work is limited in that it requires experimental measurements of bulk properties. A major future goal is to build a fully predictive model that correlates interphase properties with material characteristics and processing conditions. One pathway toward this goal entails collecting experimental data from NanoMine that consists of various materials with different combinations of fillers, polymers, and processing conditions. This inverse method can be applied as a tool to extract the interphase properties inversely, which will provide a interphase data library with different material and processing. The surface energy terms can be applied to quantify the combination of different material and processing. Then heuristic and data analytics can be applied to develop predictive tools that work across different material systems.

Investigating structure-property relationships using FEA model with gradient interphase

In Chapter 4, we proposed a multi-task, deep-learning model to predict structure-property relationships. The current construction using an FEA model with uniform interphase and the deep learning model is only trained using two-phase images, ignoring the interphase. Because of the feature-engineering-free characteristic, this method can be applied as a generalized model to build structure-property linkage for other material systems and properties, e.g. metamaterials. Although the presented work primarily relied on computational data, similar studies can be accomplished in the future if more experimental data are available. Future studies can also expand on this work by training the deep learning model using multi-phase images including interphase layers with gradient properties. For multi-phase images with gradient interphase properties, I believe that deep

learning model is even more powerful compared with traditional methods using hand crafted features. This is because for traditional methods, microstructures are only characterized by geometric and dispersion information of fillers, which will result into huge information loss if the gradient interphase profile is not considered. On the other hand, deep learning model is able to fully capture the microstructure information including the gradient profile if the gradient information is labelled using different numerical values. More importantly, the trained model can be applied to extract other material insights including the impact of gradient interphase layers and help to better understand the role of interphase in polymer nanocomposites.

Extension of the NLP toolkit to extract other information from literature sources

In order to improve the performance of the sentence classification model, future efforts should start by collecting more training data for the model. Additionally, the grammar rules applied to extract the experimental procedures and parameters can be further extended to capture more information contained within the continually evolving NanoMine schema.

In order to obtain more papers directly from different publishers, it is critical to collaborate with different publishers for API access such that paper collection can be done in an efficient and automated manner. In addition, this toolkit can be extended to extract other information, such as meta data, microstructure information, and even material properties, by building different NLP models applied to different sections of the publications.

Reference

1. Bryning, M.B., et al., *Very Low Conductivity Threshold in Bulk Isotropic Single-Walled Carbon Nanotube–Epoxy Composites.* Advanced materials, 2005. **17**(9): p. 1186-1191.
2. Kashiwagi, T., et al., *Nanoparticle networks reduce the flammability of polymer nanocomposites.* Nature materials, 2005. **4**(12): p. 928-933.
3. Zhang, S., et al., *Microstructure and electromechanical properties of carbon nanotube/poly (vinylidene fluoride—trifluoroethylene—chlorofluoroethylene) composites.* Advanced Materials, 2005. **17**(15): p. 1897-1901.
4. Moradi, M., J.A. Mohandesi, and D.F. Haghshenas, *Mechanical properties of the poly (vinyl alcohol) based nanocomposites at low content of surfactant wrapped graphene sheets.* Polymer, 2015. **60**: p. 207-214.
5. Hosseini, S.M. and M. Razzaghi-Kashani, *Vulcanization kinetics of nano-silica filled styrene butadiene rubber.* Polymer, 2014. **55**(24): p. 6426-6434.
6. Mirzaee, S., S.F. Shayesteh, and S. Mahdavifar, *Anisotropy investigation of cobalt ferrite nanoparticles embedded in polyvinyl alcohol matrix: a Monte Carlo study.* Polymer, 2014. **55**(16): p. 3713-3719.
7. Zare, Y. and H. Garmabi, *Attempts to simulate the modulus of polymer/carbon nanotube nanocomposites and future trends.* Polymer Reviews, 2014. **54**(3): p. 377-400.
8. Norouzi, M., Y. Zare, and P. Kiany, *Nanoparticles as effective flame retardants for natural and synthetic textile polymers: application, mechanism, and optimization.* Polymer Reviews, 2015. **55**(3): p. 531-560.
9. Tanaka, T., et al., *Proposal of a multi-core model for polymer nanocomposite dielectrics.* IEEE Transactions on Dielectrics and Electrical Insulation, 2005. **12**(4): p. 669-681.
10. Ciprari, D., K. Jacob, and R. Tannenbaum, *Characterization of polymer nanocomposite interphase and its impact on mechanical properties.* Macromolecules, 2006. **39**(19): p. 6565-6573.
11. *MATERIALS GENOME INITIATIVE at NIST. https://www.nist.gov/mgi. Accessed May 30 2019. .*
12. Campbell, C.E., U.R. Kattner, and Z.-K. Liu, *The development of phase-based property data using the CALPHAD method and infrastructure needs.* Integrating Materials and Manufacturing Innovation, 2014. **3**(1): p. 12.
13. Jain, A., et al., *Commentary: The Materials Project: A materials genome approach to accelerating materials innovation.* Apl Materials, 2013. **1**(1): p. 011002.
14. Lukas, H.L., S.G. Fries, and B. Sundman, *Computational thermodynamics: the Calphad method*. Vol. 131. 2007: Cambridge university press Cambridge.
15. Ceder, G. and K. Persson, *The Materials Project: A Materials Genome Approach*. 2010.
16. Kirklin, S., et al., *The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies.* npj Computational Materials, 2015. **1**: p. 15010.
17. Saal, J.E., et al., *Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD).* Jom, 2013. **65**(11): p. 1501-1509.

18. Otsuka, S., et al. *PoLyInfo: Polymer database for polymeric materials design*. in *Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on*. 2011. IEEE.
19. Ellis, B. and R. Smith, *Polymers: a property database*. 2008: CRC Press.
20. Jang, J.S., et al., *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, 2012. **33**(8): p. 1415-1423.
21. Yu, S., S. Yang, and M. Cho, *Multi-scale modeling of cross-linked epoxy nanocomposites.* Polymer, 2009. **50**(3): p. 945-952.
22. Tsai, J.-L. and S.-H. Tzeng, *Characterizing mechanical properties of particulate nanocomposites using micromechanical approach.* Journal of composite materials, 2008.
23. Yang, S. and M. Cho, *Scale bridging method to characterize mechanical properties of nanoparticle/polymer nanocomposites.* Applied Physics Letters, 2008. **93**(4): p. 043111.
24. Yablon, D.G., et al., *Quantitative viscoelastic mapping of polyolefin blends with contact resonance atomic force microscopy.* Macromolecules, 2012. **45**(10): p. 4363-4370.
25. Wang, Y. and T.H. Hahn, *AFM characterization of the interfacial properties of carbon fiber reinforced polymer composites subjected to hygrothermal treatments.* Composites science and technology, 2007. **67**(1): p. 92-101.
26. Brune, P.F., et al., *Direct Measurement of Rubber Interphase Stiffness.* Macromolecules, 2016. **49**(13): p. 4909-4922.
27. Bai, X., et al., *High-fidelity micro-scale modeling of the thermo-visco-plastic behavior of carbon fiber polymer matrix composites.* Composite Structures, 2015. **134**: p. 132-141.
28. Watcharotone, S., et al., *Interfacial and substrate effects on local elastic properties of polymers using coupled experiments and modeling of nanoindentation.* Advanced Engineering Materials, 2011. **13**(5): p. 400-404.
29. Qu, M., et al., *Nanoscale visualization and multiscale mechanical implications of bound rubber interphases in rubber–carbon black nanocomposites.* Soft Matter, 2011. **7**(3): p. 1066-1077.
30. Cheng, X., et al., *Characterization of local elastic modulus in confined polymer films via AFM indentation.* Macromolecular rapid communications, 2015. **36**(4): p. 391-397.
31. Zhang, M., et al., *Stiffness Gradients in Glassy Polymer Model Nanocomposites: Comparisons of Quantitative Characterization by Fluorescence Spectroscopy and Atomic Force Microscopy.* Macromolecules, 2017.
32. Colombini, D., G. Merle, and N. Alberola, *Use of mechanical modeling to study multiphase polymeric materials.* Macromolecules, 2001. **34**(17): p. 5916-5926.
33. Liu, H. and L.C. Brinson, *A hybrid numerical-analytical method for modeling the viscoelastic properties of polymer nanocomposites.* Journal of applied mechanics, 2006. **73**(5): p. 758-768.
34. Brune, D.A. and J. Bicerano, *Micromechanics of nanocomposites: comparison of tensile and compressive elastic moduli, and prediction of effects of incomplete exfoliation and imperfect alignment on modulus.* Polymer, 2002. **43**(2): p. 369-387.
35. Frogley, M.D., D. Ravich, and H.D. Wagner, *Mechanical properties of carbon nanoparticle-reinforced elastomers.* Composites Science and technology, 2003. **63**(11): p. 1647-1654.

36. Todd, M.G. and F.G. Shi, *Validation of a novel dielectric constant simulation model and the determination of its physical parameters.* Microelectronics journal, 2002. **33**(8): p. 627-632.
37. Qiao, R. and L.C. Brinson, *Simulation of interphase percolation and gradients in polymer nanocomposites.* Composites Science and Technology, 2009. **69**(3): p. 491-499.
38. Huang, Y., 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. IEEE.
39. Maity, P., et al., *On the size and dielectric properties of the interphase in epoxy-alumina nanocomposite.* IEEE Transactions on Dielectrics and Electrical Insulation, 2010. **17**(6): p. 1665-1675.
40. Vazquez, M., et al., *Text mining for drugs and chemical compounds: methods, tools and applications.* Molecular Informatics, 2011. **30**(6-7): p. 506-519.
41. Kim, E., et al., *Materials synthesis insights from scientific literature via text extraction and machine learning.* Chemistry of Materials, 2017. **29**(21): p. 9436-9444.
42. Kim, E., et al., *Machine-learned and codified synthesis parameters of oxide materials.* Scientific data, 2017. **4**: p. 170127.
43. Jessop, D.M., et al., *OSCAR4: a flexible architecture for chemical text-mining.* Journal of cheminformatics, 2011. **3**(1): p. 41.
44. Swain, M.C. and J.M. Cole, *ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature.* Journal of chemical information and modeling, 2016. **56**(10): p. 1894-1904.
45. Hawizy, L., et al., *ChemicalTagger: A tool for semantic text-mining in chemistry.* Journal of cheminformatics, 2011. **3**(1): p. 17.
46. Zhao, H., et al., *Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design.* APL Materials, 2016. **4**(5): p. 053204.
47. Zhao, H., et al., *NanoMine schema: An extensible data representation for polymer nanocomposites.* APL Materials, 2018. **6**(11): p. 111108.
48. Deng, H., et al., *Utilizing real and statistically reconstructed microstructures for the viscoelastic modeling of polymer nanocomposites.* Composites Science and Technology, 2012. **72**(14): p. 1725-1732.
49. Li, X., et al., *Rethinking Interphase Representations for Modeling Viscoelastic Properties for Polymer Nanocomposites.* Materialia, 2019: p. 100277.
50. Daniel, I.M., et al., *Engineering mechanics of composite materials*. Vol. 3. 1994: Oxford university press New York.
51. Bruggeman, D., *Calculation of various physics constants in heterogenous substances I Dielectricity constants and conductivity of mixed bodies from isotropic substances.* Ann. Phys, 1935. **24**(7): p. 636-664.
52. Lichtenecker, K., *Dielectric constant of natural and synthetic mixtures.* Phys. Z, 1926. **27**: p. 115.
53. Qiao, R., et al., *Effect of particle agglomeration and interphase on the glass transition temperature of polymer nanocomposites.* Journal of Polymer Science Part B: Polymer Physics, 2011. **49**(10): p. 740-748.

54. Wang, Y., et al., *Identifying interphase properties in polymer nanocomposites using adaptive optimization.* Composites Science and Technology, 2018. **162**: p. 146-155.
55. Wood, C.D., et al., *Measuring interphase stiffening effects in styrene-based polymeric thin films.* Polymer, 2015. **75**: p. 161-167.
56. Huang, Y., et al. *Modeling of charge transport in nanodielectrics using a coupled finite element and Monte Carlo approach*. in *2016 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP)*. 2016. IEEE.
57. Eslami, H., M. Rahimi, and F. Müller-Plathe, *Molecular dynamics simulation of a silica nanoparticle in oligomeric poly (methyl methacrylate): a model system for studying the interphase thickness in a polymer–nanocomposite via different properties.* Macromolecules, 2013. **46**(21): p. 8680-8692.
58. Ghanbari, A., et al., *Interphase structure in silica–polystyrene nanocomposites: a coarse-grained molecular dynamics study.* Macromolecules, 2011. **45**(1): p. 572-584.
59. Xia, W., et al., *Understanding the interfacial mechanical response of nanoscale polymer thin films via nanoindentation.* Macromolecules, 2016. **49**(10): p. 3810-3817.
60. Breneman, C.M., et al., *Stalking the Materials Genome: A Data-Driven Approach to the Virtual Design of Nanostructured Polymers.* Advanced functional materials, 2013. **23**(46): p. 5746-5752.
61. Bessa, M., et al., *A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality.* Computer Methods in Applied Mechanics and Engineering, 2017. **320**: p. 633-667.
62. Yang, Z., et al., *Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets.* Computational Materials Science, 2018. **151**: p. 278-287.
63. Hassinger, I., et al., *Toward the development of a quantitative tool for predicting dispersion of nanocomposites under non-equilibrium processing conditions.* Journal of materials science, 2016. **51**(9): p. 4238-4249.
64. Xu, H., et al., *A descriptor-based design methodology for developing heterogeneous microstructural materials system.* Journal of Mechanical Design, 2014. **136**(5): p. 051007.
65. Xu, H., et al., *A machine learning-based design representation method for designing heterogeneous microstructures.* Journal of Mechanical Design, 2015. **137**(5): p. 051403.
66. Gupta, A., et al., *Structure–property linkages using a data science approach: application to a non-metallic inclusion/steel composite system.* Acta Materialia, 2015. **91**: p. 239-254.
67. Kalidindi, S.R., *Computationally efficient, fully coupled multiscale modeling of materials phenomena using calibrated localization linkages.* ISRN Materials Science, 2012. **2012**.
68. Fast, T. and S.R. Kalidindi, *Formulation and calibration of higher-order elastic localization relationships using the MKS approach.* Acta Materialia, 2011. **59**(11): p. 4595-4605.
69. Fast, A.N., *Developing higher-order materials knowledge systems*. 2011: Drexel University.
70. Brough, D.B., D. Wheeler, and S.R. Kalidindi, *Materials knowledge systems in python—a data science framework for accelerated development of hierarchical materials.* Integrating materials and manufacturing innovation, 2017. **6**(1): p. 36-53.

71. Paulson, N.H., et al., *Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics.* Acta Materialia, 2017. **129**: p. 428-438.
72. Bansal, A., et al., *Quantitative equivalence between polymer nanocomposites and thin polymer films.* Nature materials, 2005. **4**(9): p. 693.
73. Peng, S., et al., *Local dielectric property detection of the interface between nanoparticle and polymer in nanocomposite dielectrics.* Scientific reports, 2016. **6**: p. 38978.
74. Zhang, M., et al., *Determination of Mechanical Properties of Polymer Interphase Using Combined Atomic Force Microscope (AFM) Experiments and Finite Element Simulations.* Macromolecules, 2018. **51**(20): p. 8229-8240.
75. Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.
76. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436.
77. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.
78. Kim, Y., *Convolutional neural networks for sentence classification.* arXiv preprint arXiv:1408.5882, 2014.
79. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
80. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
81. Mikolov, T., et al. *Recurrent neural network based language model*. in *Eleventh annual conference of the international speech communication association*. 2010.
82. Mikolov, T., et al. *Extensions of recurrent neural network language model*. in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011. IEEE.
83. Sak, H., A. Senior, and F. Beaufays. *Long short-term memory recurrent neural network architectures for large scale acoustic modeling*. in *Fifteenth annual conference of the international speech communication association*. 2014.
84. Yu, S., et al., *Characterization and design of functional quasi-random nanostructured materials using spectral density function.* Journal of Mechanical Design, 2017. **139**(7): p. 071401.
85. Feyel, F. and J.-L. Chaboche, *FE 2 multiscale approach for modelling the elastoviscoplastic behaviour of long fibre SiC/Ti composite materials.* Computer methods in applied mechanics and engineering, 2000. **183**(3): p. 309-330.
86. Wood, C.D., et al., *Understanding competing mechanisms for glass transition changes in filled elastomers.* Composites Science and Technology, 2016. **127**: p. 88-94.
87. Huang, Y., et al., *Predicting the breakdown strength and lifetime of nanocomposites using a multi-scale modeling approach.* Journal of Applied Physics, 2017. **122**(6): p. 065101.
88. Stein, M.L., *Interpolation of spatial data: some theory for kriging*. 2012: Springer Science & Business Media.
89. Picheny, V., T. Wagner, and D. Ginsbourger, *A benchmark of kriging-based infill criteria for noisy optimization.* Structural and Multidisciplinary Optimization, 2013. **48**(3): p. 607-626.

90. Niblack, W., *An introduction to digital image processing*. 1985: Strandberg Publishing Company.
91. Jones, D.R., M. Schonlau, and W.J. Welch, *Efficient global optimization of expensive black-box functions.* Journal of Global optimization, 1998. **13**(4): p. 455-492.
92. Jin, R., W. Chen, and A. Sudjianto. *On sequential sampling for global metamodeling in engineering design*. in *ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2002. American Society of Mechanical Engineers.
93. Balachandran, P.V., et al., *Adaptive Strategies for Materials Design using Uncertainties.* Scientific reports, 2016. **6**.
94. Sakata, S. and F. Ashida, *Ns-kriging based microstructural optimization applied to minimizing stochastic variation of homogenized elasticity of fiber reinforced composites.* Structural and Multidisciplinary Optimization, 2009. **38**(5): p. 443-453.
95. Rasmussen, C.E., *Gaussian processes for machine learning.* 2006.
96. Jones, D.R., *A taxonomy of global optimization methods based on response surfaces.* Journal of global optimization, 2001. **21**(4): p. 345-383.
97. Virtanen, S., et al., *Dielectric breakdown strength of epoxy bimodal-polymer-brush-grafted core functionalized silica nanocomposites.* IEEE Transactions on Dielectrics and Electrical Insulation, 2014. **21**(2): p. 563-570.
98. Xu, H., et al., *Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials.* Computational Materials Science, 2014. **85**: p. 206-216.
99. Park, J.-S., *Optimal Latin-hypercube designs for computer experiments.* Journal of statistical planning and inference, 1994. **39**(1): p. 95-111.
100. Schadler, L., L. Brinson, and W. Sawyer, *Polymer nanocomposites: a small part of the story.* JOM Journal of the Minerals, Metals and Materials Society, 2007. **59**(3): p. 53-60.
101. Olson, G.B., *Computational design of hierarchically structured materials.* Science, 1997. **277**(5330): p. 1237-1242.
102. Li, X., et al. *A deep adversarial learning methodology for designing microstructural material systems*. in *ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. 2018. American Society of Mechanical Engineers.
103. Li, X., et al., *A transfer learning approach for microstructure reconstruction and structure-property predictions.* Scientific reports, 2018. **8**.
104. *Dublin Core Metadata Element Set*. [cited 2018 05/29]; Available from: http://dublincore.org/documents/dces/.
105. Cecen, A., et al., *Material structure-property linkages using three-dimensional convolutional neural networks.* Acta Materialia, 2018. **146**: p. 76-84.
106. Kondo, R., et al., *Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics.* Acta Materialia, 2017. **141**: p. 29-38.
107. Iyer, A., et al., *Data-Centric Mixed-Variable Bayesian Optimization For Materials Design.* arXiv preprint arXiv:1907.02577, 2019.
108. Shen, Y., et al., *High dielectric performance of polymer composite films induced by a percolating interparticle barrier layer.* Advanced Materials, 2007. **19**(10): p. 1418-1422.

109. Ruder, S., *An overview of multi-task learning in deep neural networks.* arXiv preprint arXiv:1706.05098, 2017.
110. Yang, Z., et al., *Establishing structure-property localization linkages for elastic deformation of three-dimensional high contrast composites using deep learning approaches.* Acta Materialia, 2019. **166**: p. 335-345.
111. Zhang, Y., et al., *Microstructure reconstruction and structural equation modeling for computational design of nanodielectrics.* Integrating Materials and Manufacturing Innovation, 2015. **4**(1): p. 14.
112. Manning, C.D., C.D. Manning, and H. Schütze, *Foundations of statistical natural language processing*. 1999: MIT press.
113. Collobert, R., et al., *Natural language processing (almost) from scratch.* Journal of machine learning research, 2011. **12**(Aug): p. 2493-2537.
114. Bird, S., E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. 2009: " O'Reilly Media, Inc.".
115. Cambria, E. and B. White, *Jumping NLP curves: A review of natural language processing research.* IEEE Computational intelligence magazine, 2014. **9**(2): p. 48-57.
116. Hirschberg, J. and C.D. Manning, *Advances in natural language processing.* Science, 2015. **349**(6245): p. 261-266.
117. *DOI crawler*. [cited 2018 08/10]; Available from: https://github.com/Duke-MatSci/doi-crawler.
118. Bostanabad, R., et al., *Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques.* Progress in Materials Science, 2018.
119. *Nanomine Schema*. [cited 2018 06/24]; Available from: https://github.com/Duke-MatSci/nanomine-schema/tree/master/xml.
120. *Nanomine Homepage*. [cited 2018 06/24]; Available from: http://www.nanomine.org.
121. Dima, A., et al., *Informatics infrastructure for the materials genome initiative.* Jom, 2016. **68**(8): p. 2053-2064.
122. Zhang, Y., R. Jin, and Z.-H. Zhou, *Understanding bag-of-words model: a statistical framework.* International Journal of Machine Learning and Cybernetics, 2010. **1**(1-4): p. 43-52.
123. Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.
124. Pennington, J., R. Socher, and C. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
125. Joachims, T., *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. 1996, Carnegie-mellon univ pittsburgh pa dept of computer science.
126. Zhao, H., et al., *NanoMine Schema: A Data Representation for Polymer Nanocomposites, submitted to APL Materials, 2018.*
127. Yang, Z., et al. *Hierarchical attention networks for document classification*. in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016.

128. Cai, R., X. Zhang, and H. Wang. *Bidirectional recurrent convolutional neural network for relation classification*. in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016.
129. Krallinger, M., et al., *The CHEMDNER corpus of chemicals and drugs and its annotation principles.* Journal of cheminformatics, 2015. **7**(1): p. S2.