Research
Materials Genome Engineering—Article

Data Centric Design: A New Approach to Design of Microstructural Material Systems

Wei Chen [a],*, Akshay Iyer [a], Ramin Bostanabad [b]

[a] Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA
[b] Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA 92697, USA

ARTICLE INFO

ABSTRACT

Building processing, structure, and property (PSP) relations for computational materials design is at the heart of the Materials Genome Initiative in the era of high-throughput computational materials science. Recent technological advancements in data acquisition and storage, microstructure characterization and reconstruction (MCR), machine learning (ML), materials modeling and simulation, data processing, manufacturing, and experimentation have significantly advanced researchers' abilities in building PSP relations and inverse material design. In this article, we examine these advancements from the perspective of design research. In particular, we introduce a data-centric approach whose fundamental aspects fall into three categories: design representation, design evaluation, and design synthesis. Developments in each of these aspects are guided by and benefit from domain knowledge. Hence, for each aspect, we present a wide range of computational methods whose integration realizes data-centric materials discovery and design.

1. Introduction

Computational materials science provides a platform to achieve a deeper understanding of materials behavior across different length scales. This advancement is of particular interest to various industrial sectors, as it enables the cost-effective design of materials with engineered properties. The significance of computational materials science is also highlighted by the Materials Genome Initiative [1–4] and by the emergence of tools and frameworks such as materials by design [5,6], microstructure-sensitive design [7], and integrated computational materials engineering [8]. Since a material's morphology heavily affects its properties [9,10], the central theme of these frameworks is inverse materials design, where the link between processing, structure, and property (PSP), also known as PSP relations, is elucidated in order to engineer materials with unprecedented properties [5,11]. The non-uniqueness of inverse PSP relations, while providing design flexibility, challenges the forward development of PSP maps (Fig. 1(a)).

For most of the 20th century, materials science research and development relied on the expensive and time-consuming Edisonian approach, which involves many trials and errors. This reliance delayed the deployment of emerging materials in commercial applications. To achieve a quantum leap in materials design, we need to shift the focus of materials research from simply explaining observed phenomena to developing scientific and predictive models that explain and predict materials behavior with quantitative factors that can be controlled in order to meet the desired objectives of industrial applications. To this end, the so-called high-throughput computational materials science [12] has been developed (Fig. 1(b)). Here, the central concept is to first create a massive database that stores microstructural characteristics and properties of materials. Then, this dataset is used to train a machine learning (ML) model that can predict (or assist in the prediction of) PSP relations.

A holistic design strategy for the bi-directional traversal of PSP relations relies on addressing some key challenges: cost-effective processing techniques, microstructure representation and reconstruction, dimensionality reduction, and tractable optimization methods. The emergence of open-source materials databases [13–17] and the recent technological advancements in ML techniques [18] are accelerating our ability to address some of these challenges using a data-centric approach for materials design (Fig. 2). From the perspective of design research, the fundamental aspects
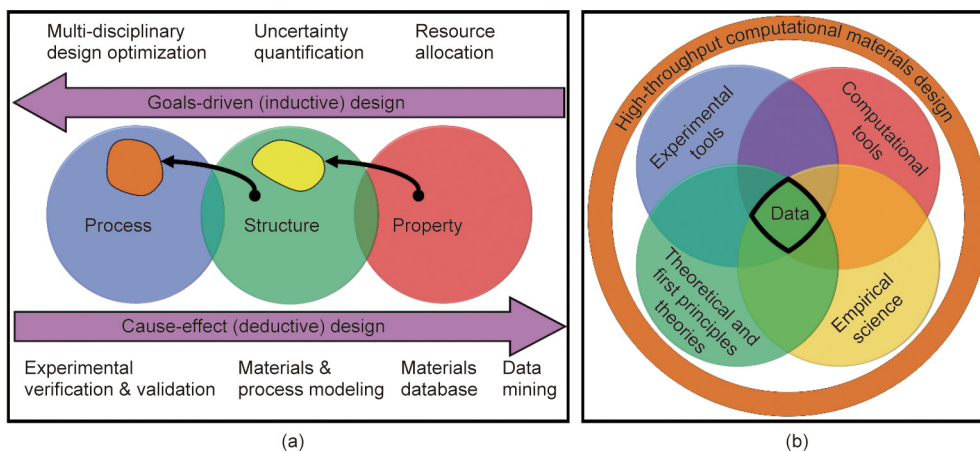
Fig. 1. (a) Forward and inverse PSP links in the Materials Genome Initiative is not unique; (b) data-driven materials design via high-throughput simulations and experiments.
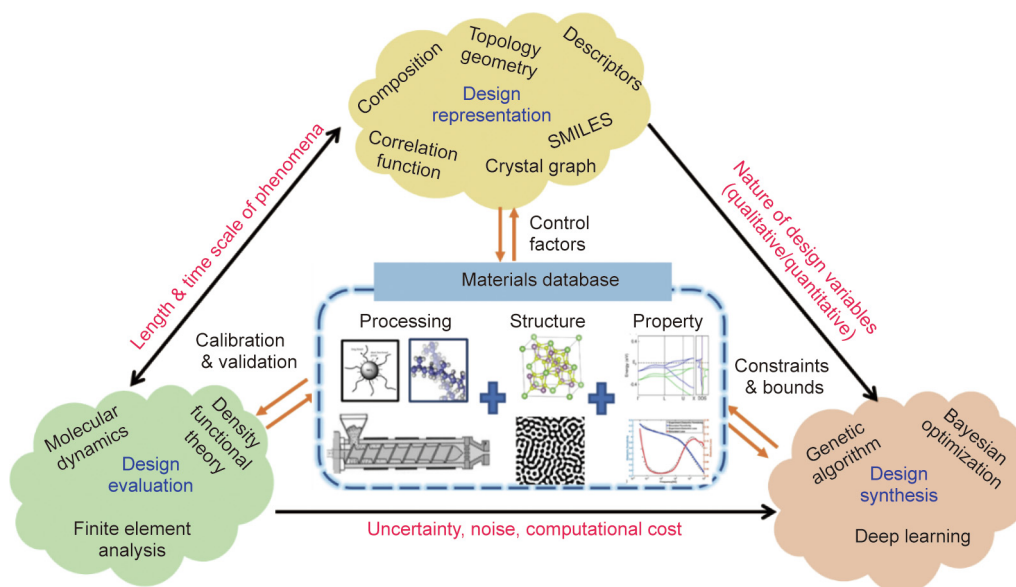

Fig. 2. Data-centric framework for materials design. SMILES: simplified molecular-input line-entry system.

of this approach fall into the categories of design representation, design evaluation, and design synthesis. Each of these aspects is guided by the knowledge gained from the PSP data stored in databases.

- **Design representation.** This encompasses methods that characterize the control factors in design—that is, the variables that influence materials' behavior. These factors depend on the material system; hence, domain knowledge can greatly help their identification. For example, the band gap of inorganic compounds is entirely determined by the composition; thus, composition is itself a suitable representation. As another example, the electrical properties of polymer nanocomposites depend on composition and microstructure. Since these two factors are high dimensional, microstructure representation methods such as spectral density function (SDF) or physical descriptors must be used for dimensionality reduction.

- **Design evaluation.** This comprises the methodologies that are employed to evaluate PSP relations. The chosen method heavily depends on both the material and the spatiotemporal scales at which the underlying phenomenon takes place. For example, density functional theory (DFT) [19,20] calculations capture atomic-level properties such as band gap; molecular

dynamics (MD) simulations model an ensemble of molecules [21–23]; and continuum mechanics is suitable for phenomena occurring at higher length scales. Each of these methods require the calibration of embedded parameters and the validation of property predictions, which is accomplished through experimental data contained in the database. ML approaches, trained on experimental data or simulated data, have been widely used to build surrogate models that replace expensive physics-based simulations.

- **Design synthesis.** This involves searching the design space to identify (feasible) optimal designs that meet the targeted properties. The choice of optimization method depends on the nature of the design variables—whether there are qualitative and quantitative design variables, the presence of uncertainty or noise in property evaluations, and the computational cost of the method. To account for manufacturing feasibility and consistency with fundamental laws and known material behaviors, constraints and bounds are often imposed during optimization to ensure feasibility.

It should be noted that the aforementioned three aspects are interrelated, as marked in Fig. 2. For example, the choice of design representation—whether mixed variable (both qualitative or quantitative) or quantitative only—will impact the choice of ML

technique in design evaluation and the choice of search algorithm in design synthesis. In this article, after first providing an overview of the role of data resources, we will review the challenges and state-of-the-art methods under each of these three aspects.

2. Materials data resources

Recent years have seen a rapid expansion of efforts toward building large data resources to accelerate materials discovery and design. The majority of such data resources are focused on metallic material systems and computational materials data, where software prediction tools can rapidly sweep through compositional space to predict specific structures and properties of interest. Examples of these data resources can be found in a recent perspective article [24]. We have been involved in developing a data resource for the design of soft materials in the field of polymer nanocomposites, called NanoMine [13,14,25] (Fig. 3). NanoMine has in-built data curation, exploration, visualization, and analysis capabilities, with curated data on over 2500 samples from the literature and individual laboratories. In principle, NanoMine offers a findable, accessible, interoperable, and reusable (FAIR) platform in which the data published in papers become directly findable and accessible via simple search tools, with open metadata standards that are interoperable with larger materials data registries; the platform also allows the easy reuse of data, such as benchmarking against new results.

At the core of developing a materials data resource is the creation of a data schema particularized for the domain of interest. The materials vocabulary used to organize the metadata framework for NanoMine forms part of the high-level "polymer data core" [26] and is compatible with the indexing from other data stores such as the materials data facility (MDF) [27,28]. Built on the MDF, we have developed an ontology-enabled knowledge graph framework [14] that helps NanoMine establish relationships between the data that falls into the following six categories:

- **Data resource.** The data in this category are the metadata of the source of the literature guided by Dublin core standards, which includes the digital object identifier (DOI) of the cited source, the authors, title, keywords, time, and source of the publication.
- **Materials.** The data in this category involve material constituent information, including the filler particle, polymer matrix, and surface treatments. The characteristics of pure matrix and filler, such as the polymer chemical structure, molecular weight, and particle density, can be entered along with the compositions (i.e., volume/weight fraction).

Fig. 3. NanoMine: an online data resource for polymer nanocomposites (www. materialsmine.org).

- **Processing.** The data in this category are a sequential description of chemical syntheses and experimental procedures. The current template provides three major categories: solution processing, melt mixing, and *in situ* polymerization. For each processing step, detailed information such as temperature, pressure, and time can be entered.
- **Characterization.** The data in this category provide information on the material characterization equipment, methods, and condition used. This information includes details on common microscopic imaging (scanning electron microscopy and transmission electron microscopy), thermal mechanical and electrical measurement, and nanoscale spectroscopy.
- **Properties.** The data in this category are measured data of material properties, including mechanical, electrical, thermal, and volumetric properties. The property data can be in the format of a scalar or a higher dimension such as two-dimensional (2D) spectroscopy or three-dimensional (3D) maps.
- **Microstructure.** The data in this category comprise raw microscopic grayscale images capturing the nanophase dispersion state. Geometric descriptors can also be included to describe the statistical characteristics of the microstructure.

The NanoMine ontology serves as an extensible knowledge-representation platform for materials science and allows the tools we develop for search, visualization, and data sharing to extend across multiple domains and interoperate with existing standards for scientific metadata. In addition to physical data, a collection of modular tools (for microstructure characterization and reconstruction (MCR) and for simulation software to model bulk nanocomposite material response) augment the knowledge generated by experiments. Integrating these different data sources to create new knowledge is critical for materials design. However, generating experimental or simulated data for the vast design space defined by the infinite combinations of constituents, microstructure morphology, and processing conditions is impractical. This signifies the need for data-centric methodologies that can effectively interrogate existing data and interpolate between them to support design representation, design evaluation, and design synthesis in discovering new high-performing materials.

3. Design representation: Microstructure characterization and reconstruction

Due to the high dimensionality of material microstructure, in microstructure-mediated design, microstructure representation is critical to ensure tractable design strategies. A good microstructure representation will ① provide significant dimension reduction; ② embody salient morphological features; ③ be physically meaningful in a way that can be easily mapped to the processing conditions; and ④ provide a computationally efficient reconstruction procedure so that statistically equivalent microstructures can be created for assessing structure–property relations and quantifying the uncertainty associated with materials heterogeneity.

MCR, coupled with ML and materials modeling and simulation, is an important component in discovering PSP relations and inverse material design in the era of high-throughput computational materials science. Given the vast diversity of microstructures observed in engineered materials, developing an MCR technique that is universally applicable is challenging. In our review article [29], we provide a comprehensive review of a wide range of MCR techniques and elaborate on their algorithmic details, their computational costs, and how they fit into the PSP mapping problems. Therein, interested readers may find detailed descriptions of multiple categories of MCR methods relying on statistical functions (such as n-point correlation functions), physical descriptors, SDF, texture synthesis, and supervised/unsupervised learning.
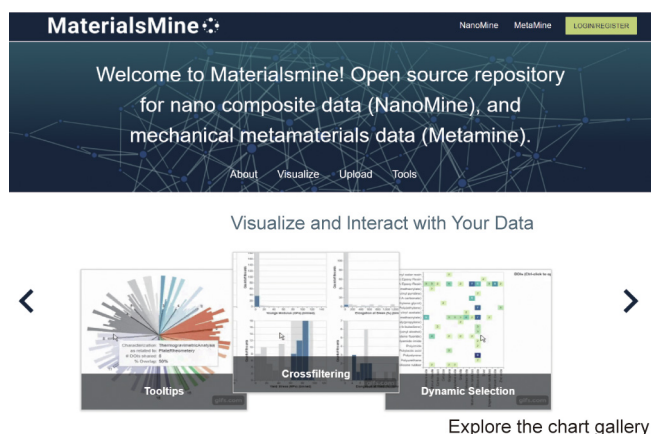
Sample MCR techniques applied to heterogeneous microstructures are illustrated in Fig. 4. Perhaps the most well-known MCR method is based on spatial correlation functions (Fig. 4(b)) [30,31], which provide probabilistic representations of the morphology but rely on a computationally intensive simulated annealing (SA) algorithm for reconstruction. The descriptor-based method (Fig. 4(a)) [32,33] represents microstructures using a small set of uncorrelated descriptors that embody significant microstructural details. Reconstruction involves a hierarchical optimization strategy to match the descriptors of reconstructed microstructures to targeted values. However, the usage of regular geometrical features and the assumption of ellipsoidal clusters deter its application for microstructures with irregular geometries. Other versions of descriptor-based MCR have been reported in the literature, as the choice of descriptor varies across materials systems and depends on the property of interest. The descriptor of nearest neighbor plays an important role in transport processes in particulate heterogeneous systems [9], microstructural evolution during recrystallization [34], particle coarsening [35], and liquid-phase sintering [34]. In fiber composites, the volume fraction (VF), size, shape, and spatial distributions of the fibers affect the mechanical properties of the composite, such as the Young's modulus, ultimate strength, and fracture toughness [36–42]. In crystalline structures, intergranular corrosion is sensitive to grain boundaries [43], so these boundaries must be used as descriptors for accurate design representation.

ML and artificial intelligence (AI) techniques, with their superior capability to learn and reconstruct complex features from isotropic/anisotropic microstructures, have gained popularity as a reconstruction tool. Applications of instance-based learning using support vector machines [44], supervised learning (Fig. 4(c)) [45,46], and transfer learning (Fig. 4(d)) [18,47] have shown good reconstruction accuracy for complex materials morphology. Transfer learning-based methods, in particular, reconstruct statistically equivalent microstructures from only one given target microstructure by leveraging a pre-trained deep convolutional neural network (CNN), Visual Geometry Group 19 (VGG-19) [48], and a loss function that measures the statistical difference between the original and the reconstructed microstructures. Knowledge obtained in the proceeding model-pruning process is then leveraged in the development of a structure–property predictive model to determine the network architecture and initialization conditions. While deep learning-based approaches are powerful for handling complex microstructure morphology, these methods often do not
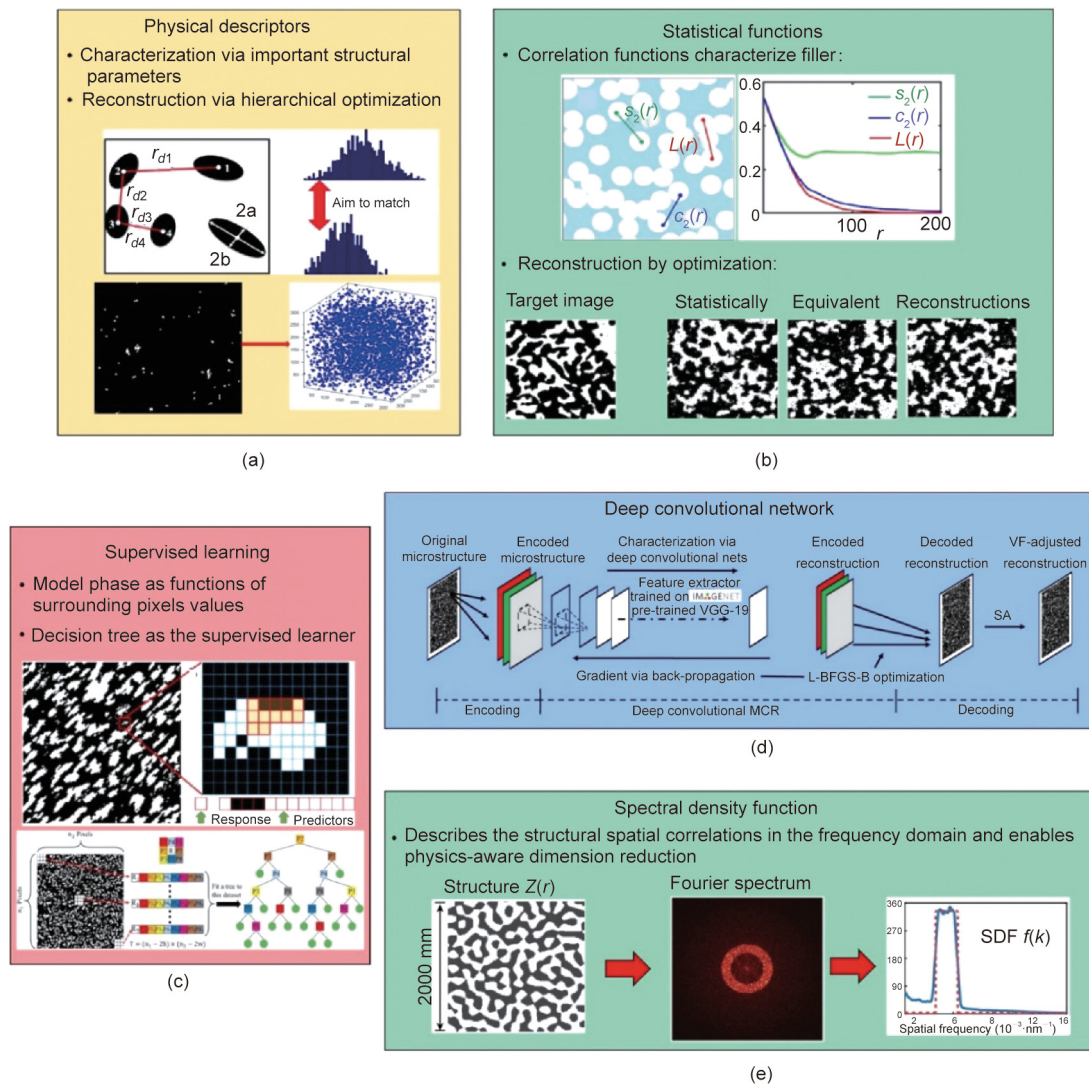

Fig. 4. Representative MCR techniques. (a) Physical descriptors; (b) statistical functions; (c) supervised learning; (d) deep convolutional network; (e) SDF. L-BFGS-B: a limited-memory quasi-Newton code for bound-constrained optimization; VGG-19: Visual Geometry Group 19, a convolutional neural network (CNN) that is 19 layers deep, trained on more than a million images from the ImageNet database.

provide the physical meaning in microstructure characterization, which hinders their use in materials design. Deep learning methods such as convolutional deep belief networks [49] and generative adversarial networks (GANs) [50] are being studied in ongoing research to provide a low-dimensional microstructure characterization that could be used as design variables.

The SDF (Fig. 4(e)) [9,51–55], a frequency domain microstructure representation, has received significant attention for its capability to provide low-dimensional and physically meaningful descriptions of quasi-random material systems with complex morphology. For isotropic materials, SDF is a one-dimensional (1D) function of spatial frequency and represents the spatial correlations in the frequency domain. Although information contained in the SDF is equivalent to a two-point autocorrelation function, Yu et al. [51] have shown that the SDF provides a more convenient representation that can be easily and sensibly mapped to both processing conditions and properties. However, the computational cost and time for reconstructing high-resolution 3D microstructures using existing approaches [56–58] remains a challenge. Moreover, while existing SDF techniques are restricted to isotropic material systems, anisotropy is highly desired in some material systems, especially where the performance is a manifestation of an underlying transport phenomenon, such as in organic photovoltaic cells (OPVCs), batteries, thermoelectric devices, and membranes for water filtration. In our recent work [59] (Fig. 5), an anisotropic microstructure design strategy that leverages the SDF for the rapid reconstruction of high-resolution, two-phase, isotropic or anisotropic microstructures in 2D and 3D is developed that quantifies anisotropy via a dimensionless scalar variable termed the anisotropy index. Application to an active layer design case study for bulk heterojunction OPVCs shows that an optimized design with strong anisotropy outperforms isotropic active-layer designs. The physics-aware SDF approach also offers significant dimension reduction in design evaluations for understanding PSP links.
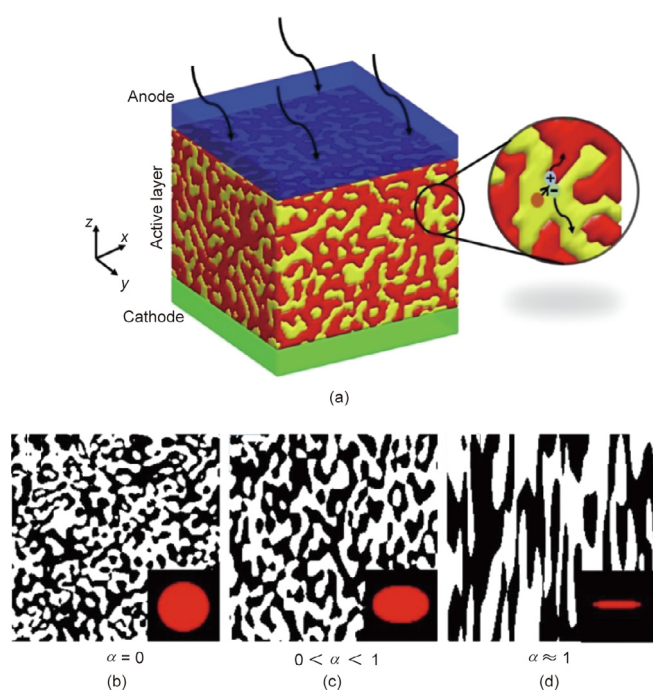
Fig. 5. (a) Schematic representation of OPVCs. Inset shows excitons (orange) dissociating into protons (blue) and electrons (green), which travel to the anode and cathode, respectively. (b–d) Quantifying anisotropy for microstructures with elliptical SDF using anisotropy index α.

4. Design evaluation: ML of PSP relations

In physics-based materials design, ML techniques have become popular surrogates for costly PSP simulators. Recent review articles on using ML and AI techniques in materials design can be found for both molecular and polymer systems [60] and metallic systems [61,62]. As shown in Fig. 6, while a wide range of statistical models such as neural networks (NNs), random forests (RFs), trees, and Gaussian processes (GPs) [63] may be considered to create surrogate models, feature identification plays a critical role in obtaining a trustworthy statistical model with good predictive capability.

The "curse of dimensionality" (i.e., the large number of descriptors or parameters) makes it extremely challenging to build predictive models with moderate sample data sizes. Hence, a combined feature-selection and feature-extraction approach is often used for dimension reduction by integrating these ML methods with materials science domain knowledge. In general, the objective of feature selection is threefold: improving predictive performance, providing more cost-effective predictors, and facilitating the discovery of underlying probabilistic principles of data generation [64]. Variable ranking is one of the most common techniques for feature selection, which enables the identification of the most informative features for building parsimonious predictive models. We have developed a range of techniques for microstructure feature selection. For example, Xu et al. [65] employed a two-step feature-selection process using descriptor pairwise correlation analysis (unsupervised learning based only on images) and the relief for regression (RReliefF) variable ranking approach [66] (supervised learning based on structure–property relations) to select the physical descriptors that best control the damping property of polymer composites. Exploratory factor analysis [67] is another technique for identifying the important features by grouping the correlated descriptors together to build a set of latent common factors. We employed factor analysis into a structural equation-modeling approach for the design of dielectric polymer composites [39]. In short, with feature selection, redundant statistical features can be dropped before further analyses are conducted.

Different from feature selection, feature extraction transforms the feature space into a lower dimensional one in which the physical interpretations are diminished. While not preserving as many physical interpretations as feature-selection methods do, feature-extraction techniques are advantageous in lowering the dimensionality of the space and are more easily trained to achieve a higher predictive accuracy [68,69]. Principle component analysis (PCA) is perhaps the most well-known linear dimensionality reduction method that can convert the high-dimensional feature space of 3D microstructure images to lower dimensional approximations [70]. It has also been demonstrated that PCA can effectively reduce the dimensionality of a two-point correlation function (commonly used in microstructure characterization) to only a few parameters [71–73]. Recent years have seen the rapid utilization of nonlinear embedding methods for feature extraction in materials design due to advances in ML techniques. One set is the bottom-up approach, in which it is assumed that a nonlinear manifold (embedded in the original feature space) governs the data distribution [74,75]. The second major set is the top–down approach, which attempts to preserve the geometric relations at all scales [76].

A wide range of ML techniques can be chosen for building a statistical model considering multiple factors, such as ① the nature of physical behavior (nonlinearity and irregularity); ② the type of input variables (qualitative, quantitative, or mixed); ③ the response of interest (continuous or classification); ④ the data source (an experiment with noise, deterministic simulation, or stochastic simulation); and ⑤ the amount of data (big or small data).
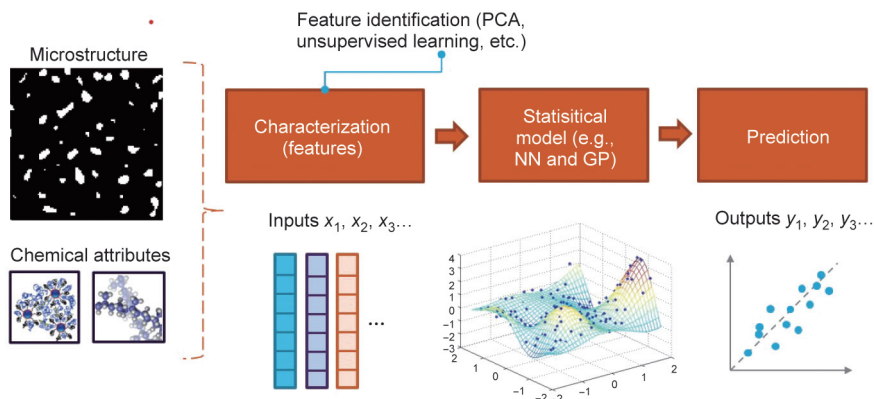
Fig. 6. Feature identification and ML in materials engineering. PCA: principle component analysis.

Due to the need to understand causal relations in PSP mappings, supervised learning methods are commonly used. While linear regression is the most straightforward approach to apply and interpret the results, methods such as decision trees [77], k-nearest neighbors (k-NNs) [78], support vector machines [79,80], and RFs [81] are better suited for more complex behavior and mixed-variable inputs; they are also flexible for creating both regression and classification models.

Recent research at the interface of ML and materials engineering has exponentially grown as big materials data are becoming increasingly available. NNs are networks connected by layers of artificial neurons, mimicking a human brain. A single neuron outputs weighted inputs through a so-called activation function. Deep neural networks (DNNs) are special NNs with more than one hidden layer that have superior learning power. For inorganic materials, crystal graph CNNs [82] have been used to model highly nonlinear behaviors using DFT-calculated thermodynamic stability entries taken from the Open Quantum Materials Database (OQMD) for accelerated materials discovery [83]. For nanocomposites, we have demonstrated that, while CNNs provide the capability of microstructure reconstruction and structure–property learning [47], GANs can be trained to learn the mapping between latent variables (LVs) and microstructures [50]. Thereafter, the low-dimensional LVs serve as design variables, and a Bayesian optimization (BO) framework can be applied to obtain microstructures with the desired material properties. For organic materials, the simplified molecular-input line-entry system (SMILES) [84] provides a meaningful representation for large molecules and has been used to design synthetic molecules using variational autoencoders [85] and reinforcement learning [86].

In the presence of small data, especially those from deterministic simulations such as DFT that require hours and days to compute one materials design, GPs provide a very viable approach. Fig. 7 is a 1D example of a GP model fitted to the collected data of $f(\cdot)$. At each input x, the output $f(x)$ is regarded as a normally distributed random variable, and the GP model predicts its mean and variance. The 95% prediction interval in the figure reflects the confidence bounds of the prediction [87,88].

Standard GP methods were developed under the premise that all input variables are quantitative, which does not hold in materials systems that involve both qualitative and quantitative design variables representing material compositions, microstructure morphology, and processing conditions. We recently proposed a latent variable Gaussian process (LVGP) [89] modeling method that maps the levels of the qualitative factor(s) to a set of numerical values for some underlying latent unobservable quantitative variable(s). In other words, the qualitative variables are "converted" to quantitative ones, and traditional GPs modeling can then be applied to
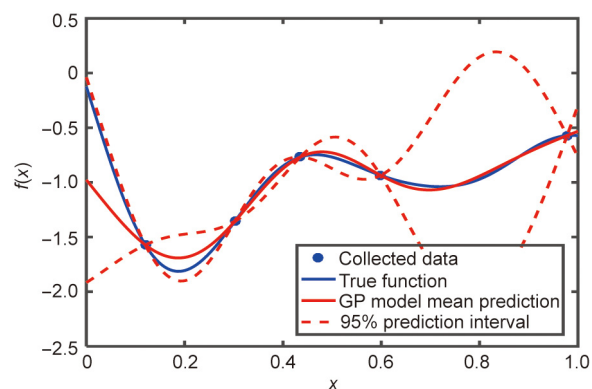

Fig. 7. A 1D example of a GP model fitted to the collected data of $f(\cdot)$.

obtain the desired model. The LV mapping of the qualitative factors provides an inherent ordering and structure for the levels of the factor(s), which leads to substantial insight into the effects of the qualitative factors. Unlike most supervised ML methods, LVGP does not require hand-crafted features to describe qualitative variables. Rather, it learns the underlying "LVs" (Z) influencing response (y) by maximizing the likelihood function.

Alleviating the need for feature engineering makes LVGP attractive for materials design applications. As conceptually illustrated in Fig. 8, the three qualitative levels of $t_j \in \{l_1, l_2, l_3\}$ of atom M in the family of M_2AX phases are associated with points in the underlying high-dimensional space of $\{v_1, v_2, \ldots\}$, defined by physical parameters such as atomic radius, ionization energy, and electron affinities. LVGP provides a nonlinear manifold mapping $z(t) = g(v_1(t), v_2(t), \ldots$ from v to the latent space Z, and the distances between the three points indicate the differences between the three levels with respect to their impact on the property of interest. The mixed-variable LVGP approach has been tested and validated for a wide range of microstructural systems such as concurrent materials selection and microstructure optimization for optimizing the light absorption of a quasi-random solar cell [90], a combinatorial search of material constitutes for optimal hybrid organic–inorganic perovskite design [90], and concurrent composition and microstructure design of nanodielectric materials [91]. Materials discovery and optimization are accomplished through the integration of the LVGP approach with BO for design synthesis, which is introduced next.

5. Design synthesis: Goal-oriented BO

Materials discovery often takes years and decades, due to several challenges associated with design synthesis: ① Even though
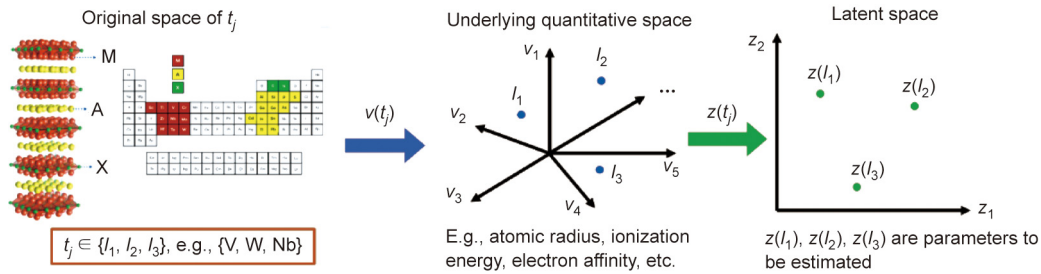
Fig. 8. Qualitative material composition selection modeled using mapping from the true high-dimensional underlying quantitative variables to the 2D LVs Z.

large datasets become available, the properties of known materials are far from the desired targets. The ML models created using the existing data are not capable of predicting behavior in the "extrapolated" regions. ② Vast combinations of candidate designs exist. In the design of organic materials, such as in polymer nanocomposite design, there are numerous choices of material constituents (e.g., the types of filler and matrix) and processing conditions (e.g., the type of surface treatment); each combination follows drastically different physical mechanisms with significant impact on the overall properties. In the design of inorganic materials such as microelectronics, the possible options of atomic structure–composition variable spaces are in the order of millions, defined by different structure prototypes (crystal graphs), composition (choice of chemistry elements), and stoichiometry (ratio of elements). ③ The existence of both quantitative and qualitative material design variables results in multiple disjoint regions in the property/performance space. The combinatorial nature poses additional challenges in materials modeling and the search for an optimal solution.

During the past half decade, the BO approach has emerged as the most effective approach to materials design synthesis [92–95], due to its capability of locating the global optima for highly nonlinear functions within tens to hundreds of objective-function (i.e., material property) evaluations. Starting from a small dataset, BO relies on an adaptive sampling technique to approach the global optimum efficiently—an attractive feature for materials design. Fig. 9 shows our proposed on-demand goal-driven data augmentation framework, integrating curated material databases with material property simulations and ML. The framework is initiated from a database of curated experimental and simulated data describing material properties with appropriate attributes. Based on PSP relationships, one identifies a subset of attributes that are known to

influence material properties and act as design variables in BO. These attributes may be quantitative (e.g., microstructure descriptors or interphase descriptors) or qualitative (e.g., type of filler, polymer, or a combination of both).

Using the predictions and uncertainty quantification of the ML model, Bayesian inference determines the design that shows the most "potential" for improvement in terms of material property. There are several metrics—commonly known as acquisition functions—for evaluating "potential" improvement. The acquisition function strikes a balance between exploration (reducing prediction uncertainty) and exploitation (optimizing the design objective) of the design space. The most commonly used acquisition functions are expected improvement (EI) [96] and probability of improvement [97]. Once a promising design is identified by the acquisition function, its corresponding material property is evaluated using "on-demand" experiments, simulations, or both. The nature of simulations depends on the material system and property under consideration, often requiring the calibration of parameters. For example, finite-element simulations for the prediction of dielectric properties in nanocomposites require the calibration of interphase-shifting parameters [98]. Once the property evaluation is complete, the design is added to the database and the above steps are repeated. The termination criterion is usually the maximum number of iterations, which depends on the cost and time required for the simulations or experiments.

By integrating the mixed-variable LVGP model introduced in Section 4 and the BO framework, we have successfully applied the BO approach to designs of organic, inorganic, and hybrid materials. For example, in concurrent composition and microstructure design [91], the design of electrically insulating nanocomposites is cast as a multicriteria optimization problem with the goal of maximizing the dielectric breakdown strength while minimizing
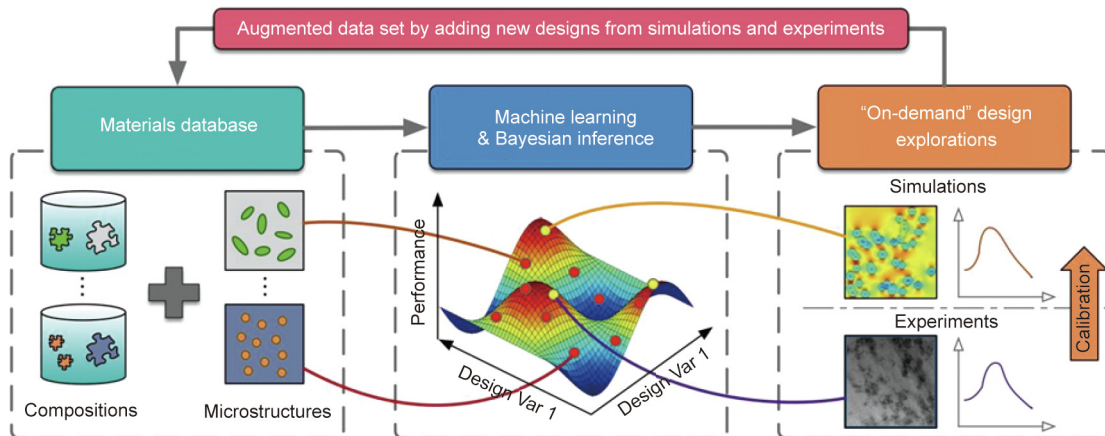

Fig. 9. The BO approach treats the existing dataset as prior knowledge, chooses new samples, and builds ML models using curated and new experimental and simulation data to capture PSP relations for optimization.

the dielectric permittivity and dielectric loss (Fig. 10). The SDF is selected as the microstructure representation, with the underlying function type identified based on experimental images. Within tens of simulations and using the multi-response LVGP approach, our method identifies a diverse set of designs on the Pareto frontier indicating the tradeoff between dielectric properties. This method was shown to be much more efficient than using genetic algorithms.

The generality of BO using LVGP is further exemplified by a combinatorial search for an ABX$_3$ hybrid organic–inorganic perovskite with optimal binding energy to solvents [90]. The design space consists of three choices each for the A and X sites and eight choices for the type of solvent, while the B site remains unchanged. In addition, the three X's can be chosen independently. Out of the 648 possible ABX$_3$-solvent combinations, 240 are stable and constitute the search space for BO. Fig. 11(a) shows that BO converges to the optimal combination faster with LVGP, as compared with the multiplicative covariance (MC) [99,100] GP model commonly used for qualitative variables hitherto. Furthermore, the latent space estimated by LVGP provides insights into the nature of the levels for each qualitative variable. The positioning of solvent choice 1 and 7 far from the others in Fig. 11(b) indicates that their effects on the binding energy are distinct. This insight is validated by analyzing the distribution of binding energies in Fig. 11(c), which shows that combinations with solvents 1 and 7 result in higher binding energies. Several materials design applications can be cast as a combinatorial optimization problem. For example, we recently demonstrated that the search for functional electronic materials

design with metal–insulator transitions (MITs) [101] can be expedited with LVGP-based multicriteria BO. These findings indicate that integrating mixed-variable LVGP models with BO is an effective approach for design synthesis in the design of engineered material systems.

6. Conclusions

Here, we presented a data-centric approach for materials design that integrates state-of-the-art computational techniques for microstructural analysis and design. These techniques fall into the categories of design representation, design evaluation, and design synthesis. Realization of this approach is supported by the creation of materials data hubs such as NanoMine, where a wide range of data resources and tools are developed for microstructural analysis and optimal materials design. As we have illustrated, this development consists of the systematic integration of image pre-processing, microstructure characterization, reconstruction, dimension reduction, ML of PSP relations, and multi-objective optimization.

A key question for achieving a seamless integration of design representation, design evaluation, and design synthesis is: What is the proper microstructure representation for the materials systems of interest? We presented a range of microstructure representation techniques based on correlation functions, physical descriptors, SDF, supervised learning, and deep learning. While the merits of these different techniques vary from one system to another, it is evident that stochasticity plays a critical role and
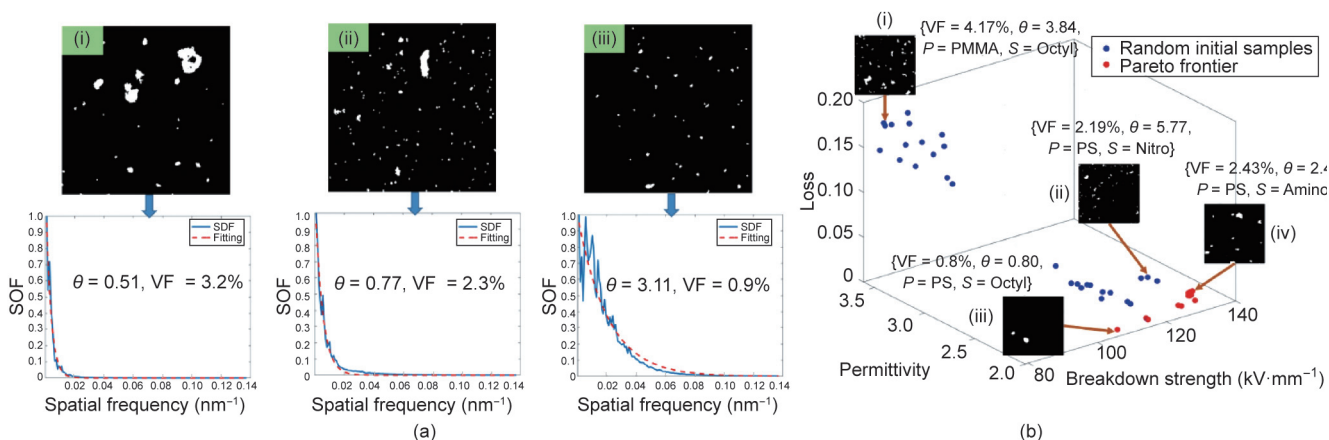
Fig. 10. Concurrent composition and microstructure design for nanocomposites. (a) SDF characterizes nanoparticle dispersion using parameter θ and nanoparticles loading using the VF. (b) Multicriteria mixed-variable BO using LVGP identifies the Pareto frontier displaying significant improvement with respect to randomly selected initialization samples (*P* stands for polymer type; *S* stands for type of surface treatment. PMMA: polymethyl methacrylate; PS: polystyrene).
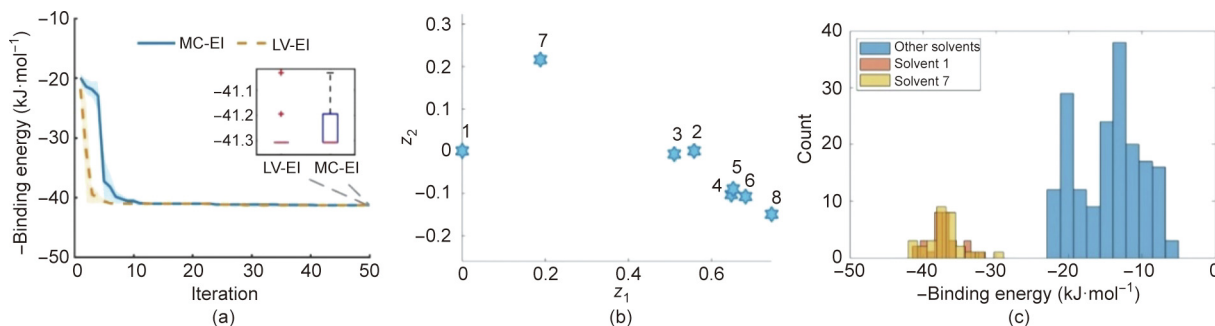
Fig. 11. (a) Comparing the convergence BO with the EI acquisition function for MC-EI and the LV-EI GP. (b) Latent space for the "solvent type" categorical variable with eight levels. (c) Distribution of binding energy categorized by "solvent type."

must be considered in materials representation and property predictions.

For design evaluation, ML approaches have played an increasingly important role in knowledge discovery and in building surrogate models that replace physics-based simulations. Since big data and lack of data co-exist in materials informatics, care must be exercised to ensure that the selected ML technique, such as NN, RF, or GP, is consistent with the data availability. As more materials data are being generated, deep learning is gaining popularity for image-based materials informatics, in which interpretation of the learned microstructural features relies on developing explainable deep models.

Finally, ML should not be viewed as an isolated component in materials discovery. For example, its integration with information-theoretic approaches such as BO can provide a significant speedup. As materials discovery is combinatorial in nature, it requires mixed-variable models such as LVGP that can handle both qualitative and quantitative design variables. These models provide quantitative measures of "distances" for different materials concepts based on their influence on the desired material properties. More research is needed to extend the current methods to handle high-dimensional materials design problems with millions or billions of combinations. The same information-theoretic framework can be extended to guide the design of batch samples and high-throughput experiments.

Acknowledgments

Compliance with ethics guidelines

Wei Chen, Akshay Iyer, and Ramin Bostanabad declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] National Science and Technology Council (US). Materials genome initiative for global competitiveness [Internet]. Washington DC: Executive Office of the President, National Science and Technology Council; 2011 Jun 24. Available from: https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf.
[2] Olson GB. Preface to the viewpoint set on: the materials genome. Scr Mater 2014;70:1–2.
[3] Ward C. Materials Genome Initiative for global competitiveness. In: Proceedings of the 23rd Advanced Aerospace Materials and Processes (AeroMat) Conference and Exposition; 2012 Jun 18–21; Charlotte, NC, USA; 2012.
[4] McDowell DL, Kalidindi SR. The materials innovation ecosystem: a key enabler for the materials genome initiative. MRS Bull 2016;41(4):326–37.
[5] Olson GB. Computational design of hierarchically structured materials. Science 1997;277(5330):1237–42.
[6] Olson GB. Designing a new material world. Science 2000;288(5468):993–8.
[7] Fullwood DT, Niezgoda SR, Adams BL, Kalidindi SR. Microstructure sensitive design for performance optimization. Prog Mater Sci 2010;55(6):477–562.
[8] Committee on Integrated Computational Materials Engineering. Integrated computational materials engineering: a transformational discipline for improved competitiveness and national security. Washington, DC: National Academies Press; 2008.
[9] Torquato S. Random heterogeneous materials: microstructure and macroscopic properties. New York: Springer-Verlag New York; 2002.
[10] Kumar H, Briant CL, Curtin WA. Using microstructure reconstruction to model mechanical behavior in complex microstructures. Mech Mater 2006;38(8–10):818–32.
[11] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater 2016;4(5):053208.
[12] Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. Nat Mater 2013;12(3):191–201.
[13] Zhao H, Li X, Zhang Y, Schadler LS, Chen W, Brinson LC. Perspective: NanoMine: a material genome approach for polymer nanocomposites analysis and design. APL Mater 2016;4(5):053204.
[14] Zhao H, Wang Y, Lin A, Hu B, Yan R, McCusker J, et al. NanoMine schema: an extensible data representation for polymer nanocomposites. APL Mater 2018;6(11):111108.
[15] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002.
[16] Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM 2013;65(11):1501–9.
[17] Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, et al. AFLOWLIB. ORG: a distributed materials properties repository from high-throughput *ab initio* calculations. Comput Mater Sci 2012;58:227–35.
[18] Bostanabad R. Reconstruction of 3D microstructures from 2D images via transfer learning. Comput Aided Des 2020;128:102906.
[19] Koch W, Holthausen MC. A chemist's guide to density functional theory. 2nd ed. Medford: John Wiley & Sons; 2015.
[20] Parr RG. Density functional theory of atoms and molecules. In: Fukui K, Pullman A, editors. Horizons of quantum chemistry. Dordrecht: Springer; 1980. p. 5–15.
[21] Duan K, He Y, Li Y, Liu J, Zhang J, Hu Y, et al. Machine-learning assisted coarse-grained model for epoxies over wide ranges of temperatures and cross-linking degrees. Mater Des 2019;183:108130.
[22] Bejagam KK, Singh S, An Y, Deshmukh SA. Machine-learned coarse-grained models. J Phys Chem Lett 2018;9(16):4667–72.
[23] Wang W, Gómez-Bombarelli R. Coarse-graining auto-encoders for molecular dynamics. npj Comput Mater 2019;5(1):1–9.
[24] Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci 2019;6(21):1900808.
[25] Brinson LC, Deagen M, Chen W, McCusker J, McGuinness DL, Schadler LS, et al. Polymer nanocomposite data: curation, frameworks, access, and potential for discovery and design. ACS Macro Lett 2020;9(8):1086–94.
[26] Therneau T, Atkinson B, Ripley B. rpart: recursive partitioning and regression trees. Version 4.1-10 [software]. 2019 May 1. Available from: https://rdrr.io/cran/rpart/.
[27] Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I. The materials data facility: data services to advance materials science research. JOM 2016;68(8):2045–52.
[28] Blaiszik B, Ward L, Schwarting M, Gaff J, Chard R, Pike D, et al. A data ecosystem to support machine learning in materials science. MRS Commun 2019;9(4):1125–33.
[29] Bostanabad R, Zhang Y, Li X, Kearney T, Brinson LC, Apley DW, et al. Computational microstructure characterization and reconstruction: review of the state-of-the-art techniques. Prog Mater Sci 2018;95:1–41.
[30] Yeong CLY, Torquato S. Reconstructing random media. Phys Rev E 1998;57(1):495–506.
[31] Yeong CLY, Torquato S. Reconstructing random media. II. Three-dimensional media from two-dimensional cuts. Phys Rev E 1998;58(1):224–33.
[32] Xu H, Dikin DA, Burkhart C, Chen W. Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials. Comput Mater Sci 2014;85:206–16.
[33] Xu H, Li Y, Brinson C, Chen W. A descriptor-based design methodology for developing heterogeneous microstructural materials system. J Mech Des 2014;136(5):051007.
[34] Snyder VA, Alkemper J, Voorhees PW. The development of spatial correlations during Ostwald ripening: a test of theory. Acta Mater 2000;48(10):2689–701.
[35] DeHoff RT. A geometrically general-theory of diffusion controlled coarsening. Acta Metall Mater 1991;39(10):2349–60.
[36] Li M, Ghosh S, Richmond O, Weiland H, Rouns TN. Three dimensional characterization and modeling of particle reinforced metal matrix composites: part I: quantitative description of microstructural morphology. Mater Sci Eng A 1999;265(1–2):153–73.
[37] Nan CW, Clarke DR. The influence of particle size and particle fracture on the elastic/plastic deformation of metal matrix composites. Acta Mater 1996;44(9):3801–11.
[38] Breneman CM, Brinson LC, Schadler LS, Natarajan B, Krein M, Wu K, et al. Stalking the materials genome: a data-driven approach to the virtual design of nanostructured polymers. Adv Funct Mater 2013;23(46):5746–52.
[39] Zhang Y, Zhao H, Hassinger I, Brinson LC, Schadler LS, Chen W. Microstructure reconstruction and structural equation modeling for computational design of nanodielectrics. Integr Mater Manuf Innov 2015;4:209–34.
[40] Karásek L, Sumita M. Characterization of dispersion state of filler and polymer-filler interactions in rubber-carbon black composites. J Mater Sci 1996;31:281–9.

[41] Yuan M, Turng LS. Microstructure and mechanical properties of microcellular injection molded polyamide-6 nanocomposites. Polymer 2005;46 (18):7273–92.

[42] Baghgar M, Barnes AM, Pentzer E, Wise AJ, Hammer BAG, Emrick T, et al. Morphology-dependent electronic properties in cross-linked (P3HT-*b*-P3MT) block copolymer nanostructures. ACS Nano 2014;8(8):8344–9.

[43] Rollett AD, Lee SB, Campman R, Rohrer GS. Three-dimensional characterization of microstructure by electron back-scatter diffraction. Annu Rev Mater Res 2007;37:627–58.

[44] Sundararaghavan V, Zabaras N. Classification and reconstruction of three-dimensional microstructures using support vector machines. Comput Mater Sci 2005;32(2):223–39.

[45] Bostanabad R, Bui AT, Xie W, Apley DW, Chen W. Stochastic microstructure characterization and reconstruction via supervised learning. Acta Mater 2016;103:89–102.

[46] Bostanabad R, Chen W, Apley DW. Characterization and reconstruction of 3D stochastic microstructures via supervised learning. J Microsc 2016;264 (3):282–97.

[47] Li X, Zhang Y, Zhao H, Burkhart C, Brinson LC, Chen W. A transfer learning approach for microstructure reconstruction and structure–property predictions. Sci Rep 2018;8(1):13461.

[48] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.

[49] Cang R, Xu Y, Chen S, Liu Y, Jiao Y, Ren MY. Microstructure representation and reconstruction of heterogeneous materials via deep belief network for computational material design. J Mech Des 2017;139(7):071404.

[50] Yang Z, Li X, Catherine Brinson L, Choudhary AN, Chen W, Agrawal A. Microstructural materials design via deep adversarial learning methodology. J Mech Des 2018;140(11):111416.

[51] Yu S, Zhang Y, Wang C, Lee WK, Dong B, Odom TW, et al. Characterization and design of functional quasi-random nanostructured materials using spectral density function. J Mech Des 2017;139(7):071401.

[52] Uche OU, Stillinger FH, Torquato S. Constraints on collective density variables: two dimensions. Phys Rev E 2004;70(4):046122.

[53] Uche OU, Torquato S, Stillinger FH. Collective coordinate control of density distributions. Phys Rev E 2006;74(3):031104.

[54] Batten RD, Stillinger FH, Torquato S. Classical disordered ground states: super-ideal gases and stealth and equi-luminous materials. J Appl Phys 2008;104(3):033504.

[55] Florescu M, Torquato S, Steinhardt PJ. Designer disordered materials with large, complete photonic band gaps. Proc Natl Acad Sci 2009;106 (49):20658–63.

[56] Cahn JW. Phase separation by spinodal decomposition in isotropic systems. J Chem Phys 1965;42(1):93–9.

[57] Teubner M. Level surfaces of Gaussian random fields and microemulsions. EPL 1991;14(5):403–8.

[58] Chen D, Torquato S. Designing disordered hyperuniform two-phase materials with novel physical properties. Acta Mater 2018;142:152–61.

[59] Iyer A, Dulal R, Zhang Y, Ghumman UF, Chien T, Balasubramanian G, et al. Designing anisotropic microstructures with spectral density function. Comput Mater Sci 2020;179:109559.

[60] Chen G, Shen Z, Iyer A, Ghumman UF, Tang S, Bi J, et al. Machine-learning-assisted *de novo* design of organic molecules and polymers: opportunities and challenges. Polymers 2020;12(1):163.

[61] Johnson NS, Vulimiri PS, To AC, Zhang X, Brice CA, Kappes BB, et al. Invited review: machine learning for materials developments in metals additive manufacturing. Addit Manuf 2020;36:101641.

[62] Bock FE, Aydin RC, Cyron CJ, Huber N, Kalidindi SR, Klusemann B. A review of the application of machine learning and data mining approaches in continuum materials mechanics. Front Mater 2019;6:110.

[63] Bostanabad R, Chan YC, Wang LW, Zhu P, Chen W. Globally approximate Gaussian processes for big data with application to data-driven metamaterials design. J Mech Des 2019;141(11):111402.

[64] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[65] Xu H, Liu R, Choudhary A, Chen W. A machine learning-based design representation method for designing heterogeneous microstructures. J Mech Des 2015;137(5):051403.

[66] Robnik-Šikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression. In: Proceedings of the Fourteenth International Conference on Machine Learning; 1997 Jul 8–12; Nashville, TN, USA. San Francisco: Morgan Kaufmann Publishers, Inc.; 1997. p. 296–304.

[67] Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4(3):272–99.

[68] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290(5500):2323–6.

[69] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23.

[70] Jolliffe IT. Principal component analysis. In: Everitt BS, Howell D, editors. Encyclopedia of statistics in behavioral science. Hoboken: John Wiley & Sons, Inc.; 2005.

[71] Yabansu YC, Steinmetz P, Hötzer J, Kalidindi SR, Nestler B. Extraction of reduced-order process-structure linkages from phase-field simulations. Acta Mater 2017;124:182–94.

[72] Popova E, Rodgers TM, Gong X, Cecen A, Madison JD, Kalidindi SR. Process-structure linkages using a data science approach: application to simulated additive manufacturing data. Integr Mater Manuf Innov 2017;6(1):54–68.

[73] Paulson NH, Priddy MW, McDowell DL, Kalidindi SR. Reduced-order structure–property linkages for polycrystalline microstructures based on 2-point statistics. Acta Mater 2017;129:428–38.

[74] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Dietterich T, Becker S, Ghahramani Z, editors. Advances in neural information processing systems. Cambridge: MIT Press; 2001. p. 585–91.

[75] Donoho DL, Grimes C. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci 2003;100 (10):5591–6.

[76] Saxena A, Gupta A, Mukerjee A. Non-linear dimensionality reduction by locally linear isomaps. In: Proceedings of the 11th International Conference on Neural Information Processing; 2004 Nov 22–25; Calcutta, India. Berlin: Springer; 2004. p. 1038–43.

[77] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. London: Taylor & Francis Group; 1984.

[78] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.

[79] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20 (3):273–97.

[80] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory; 1992 Jul 27–29; Pittsburgh, PA, USA; 1992; p. 144–152.

[81] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[82] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018;120(14):145301.

[83] Park CW, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Phys Rev Mater 2020;4(6):063801.

[84] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28(1):31–6.

[85] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4(2):268–76.

[86] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design. Sci Adv 2018;4(7):aap7885.

[87] Tao S, Shintani K, Bostanabad R, Chan YC, Yang G, Meingast H, et al. Enhanced Gaussian process metamodeling and collaborative optimization for vehicle suspension design optimization. In: Proceedings of the ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference; 2017 Aug 6–9; Cleveland, OH, USA; 2017.

[88] Bostanabad R, Kearney T, Tao S, Apley DW, Chen W. Leveraging the nugget parameter for efficient Gaussian process modeling. Int J Numer Methods Eng 2018;114(5):501–16.

[89] Zhang Y, Tao S, Chen W, Apley DW. A latent variable approach to Gaussian process modeling with qualitative and quantitative factors. Technometrics 2020;62(3):291–302.

[90] Zhang Y, Apley DW, Chen W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. Sci Rep 2020;10(1):4924.

[91] Iyer A, Zhang Y, Prasad A, Tao S, Wang Y, Schadler L, et al. Data centric mixed variable Bayesian optimization for materials design. In: Proceedings of the ASME International Design Engineering Technical Conference; 2019 Aug 18–21; Anaheim, CA, USA; 2019.

[92] Balachandran PV, Xue D, Theiler J, Hogden J, Lookman T. Adaptive strategies for materials design using uncertainties. Sci Rep 2016;6(1):19660.

[93] Li C, de Celis Leal DR, Rana S, Gupta S, Sutti A, Greenhill S, et al. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. Sci Rep 2017;7(1):5683.

[94] Yamashita T, Sato N, Kino H, Miyake T, Tsuda K, Oguchi T. Crystal structure prediction accelerated by Bayesian optimization. Phys Rev Mater 2018;2 (1):013803.

[95] Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput Mater 2019;5(1):21.

[96] Mockus J, Tiesis V, Zilinskas A. The application of Bayesian methods for seeking the extremum. In: Dixon LCW, Szego GP, editors. Towards global optimization. Amsterdam: Elsevier; 1978. p. 117–29.

[97] Kushner HJ. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. J Basic Eng 1964;86(1):97–106.

[98] Wang Y, Zhang Y, Zhao H, Li X, Huang Y, Schadler LS, et al. Identifying interphase properties in polymer nanocomposites using adaptive optimization. Compos Sci Technol 2018;162:146–55.

[99] Zhang Y, Notz WI. Computer experiments with qualitative and quantitative variables: a review and reexamination. Qual Eng 2015;27(1):2–13.

[100] McMillan NJ, Sacks J, Welch WJ, Gao F. Analysis of protein activity data by Gaussian stochastic process models. J Biopharm Stat 1999;9(1):145–60.

[101] Wang Y, Iyer A, Chen W, Rondinelli JM. Featureless adaptive optimization accelerates functional electronic materials design. Appl Phys Rev 2020;7:041403.