# Data Driven Modeling, Statistical Analysis, and Machine Learning for Additive Manufacturing

N.S. Johnson,[1] R. Liu,[1] X. Zhang,[1] C.A. Brice,[1] B. Kappes,[1] H. Wang,[2]
B. Meredig,[3] J. Ling,[3] J. Saal,[3] C.K.H. Borg,[3] and A.P. Stebner*[1]

[1] *Department of Mechanical Engineering, Colorado School of Mines, Golden, CO 80401*
[2] *Department of Computer Science, Colorado School of Mines, Golden, CO 80401*
[3] *Citrine Informatics, Redwood City, CA 94063*

In metal additive manufacturing (AM), materials and components are concurrently made in a single process as layers of metal are fabricated on top of each other in the (near) final topology required for the end-use product. Consequently, a large number of processing degrees of freedom (tens to hundreds) must be simultaneously controlled and understood; hence, metal AM is a highly interdisciplinary technology that requires synchronized consideration of physics, chemistry, materials science, physical metallurgy, computer science, electrical engineering, and mechanical engineering. The use of modern statistics-based approaches to modeling data sets with many degrees of freedom (known as machine learning) with metal AM can reduce the time and cost to elucidate and optimize the complex multidisciplinary phenomena. Machine learning techniques have been used in materials science for several decades. Most prolifically, the density functional theory community (DFT) rapidly adopted machine learning and has used it since the early 2000s for evaluating many combinations of elements and crystal structures to discover new materials. This focused review examines the potential of machine learning in metal AM, highlighting the many parallels to previous efforts in materials science and manufacturing, and discusses new challenges specific to metal AM.

## CONTENTS

I. Motivation 1

II. Phrasing Additive Manufacturing as a Machine Learning Problem 2
- A. The Design Space of Additive Manufacturing 2
- B. Additive Manufacturing Data Types and Formats 2
- C. The Assumptions Behind Machine Learning 3
- D. Unsupervised Machine Learning 4
- E. Supervised Machine Learning 5

III. Current ICME Tools are Well Equipped to Integrate with an ML Framework 5
- A. Pre-Build Design 5
  - 1. Alloy Design 5
  - 2. Design of Experiments 8
  - 3. Topology Optimization 9
- B. Process Design 10
  - 1. Model Complexity and Dimensionality Reduction 10
  - 2. Surrogate Modeling 12
- C. Process Monitoring and Characterization 14
  - 1. In Situ Process Monitoring and Feedback 14
  - 2. Featurization of Qualitative Image Data 15

IV. Learning from the Past: Moving Towards Database-Driven Design of Additive 16

V. Conclusions 18

References 18

## I. MOTIVATION

Metals-based additive manufacturing (AM) represents a potential paradigm shift in how products are manufactured, providing versatility in the type and design of parts produced by a single manufacturing facility, decentralizing manufacturing capability, and enabling novelty in the properties and design of the parts, to name but a few benefits AM offers. However, there have been significant roadblocks to fully realize AM's potential, particularly in the control of consistency and quality in part production and in the development of materials amenable to the AM process. Although decades of immense scientific and engineering work in industry, academia, and government have produced large advances in making AM a practical manufacturing solution, the metallurgical challenges facing AM persist. Computational materials simulation and the Integrated Computational Materials Engineering (ICME) approach have made strides in accelerating materials development, but the features that make AM such a departure from traditional manufacturing requires more uniquely suited problem-solving methods. In this paper, we argue that machine learning-based materials informatics is such a solution, capable of significantly accelerating the AM development process.

The 20th century saw the maturation of materials science and engineering as a field of study, enabling more rapid development of novel materials and materials manufacturing methods for specific applications. The process-structure-properties-performance paradigm transformed the combinatorial trial-and-error and intuition-driven materials discovery process into a problem of engineering a desired microstructure through designed processing. For instance, the history of tur-

bine blade superalloy development is typified by advancements in control over microstructure through processing, including increasingly complex alloying recipes, multi-step heat treatments, and single-crystal casting.

In the past decades, materials development has greatly accelerated to match the broader acceleration of general technology advancement. Computational materials science has enabled the prediction of microstructure from processing and of properties from microstructure, reducing the need for costly and time consuming experimentation. The Integrated Computational Materials Engineering (ICME) approach tightly integrates physics-based computational models into the industrial design process, allowing the desired performance requirements of a part to guide the design of a novel material. Examples include low-RE Ni superalloys for better turbine performance and lower cost and the Ferrium S53 alloy designed for corrosion-resistant landing gears. Both cases took materials development timelines from decades to years, demonstrating the practical capability of designing new materials within an industrial product timeframe. Generalizing this capability to more industries and further accelerating the process is the primary goal of the Materials Genome Initiative (MGI).

Current AM capabilities are limited by materials-based problems which are uniquely difficult to solve with the above paradigm. The AM process itself is complex relative to traditional casting methods, including rapid solidification, vaporization and ingestion of volatile elements, and a complex thermal history, all of which vary with part location and require advanced computational tools to properly predict microstructure and properties. However, the lower cost and time barrier to entry for performing AM has enabled the rapid accumulation of experimental data, enabling the Edisonian approach for finding optimized AM processing methods and parameters of existing alloys. For AM, the ICME-based tools have been catching up, attempting to bootstrap legacy models to this data, with limited success. As such, current AM materials development is largely combinatorial, as the processing of legacy alloys are optimized with extensive design of experiments and AM-specific alloys with higher performance are just now being effectively developed.

It is within this context we argue that machine learning (ML) can accelerate the application of additive manufacturing. ML as a method for model development has shown wide application in the past years, in industries ranging from finance to social networking. The use of ML in materials has been relatively limited for a variety of reasons, primarily the lack of a curated and large dataset on which ML can operate. The MGI identified this as a primary problem for accelerating materials development, and there has been significant progress recently in infrastructure development for materials databases suitable for informatics tools such as ML. The difficulty in producing physics-based ICME models and the ability to rapidly and efficiently sample processing space in a combinatorial approach makes additive manufacturing an attractive use case for ML. Machine learning as a framework can couple the legacy ICME tools with the experimental data to produce much more accurate AM process-structure-property models and to automate the iteration of designed experiments for model improvement and optimized materials.

In this paper, we present our arguments for the use of ML to address AM challenges. We begin by detailing how ML could be applied to AM and the use cases we envision. Then we discuss existing AM models and how they could be integrated into a ML framework. Existing examples of ML as applied to materials and AM specifically are then reviewed. We conclude with a prospective outlook on the potential of ML for AM.

## II. PHRASING ADDITIVE MANUFACTURING AS A MACHINE LEARNING PROBLEM

### A. The Design Space of Additive Manufacturing

The **design space** of additive manufacturing is the set of all machine parameters and measurable manufacturing phenomenon or material properties that result from the manufacturing process. An example of controllable machine parameters can be found in Table I. Measurable manufacturing phenomenon include quantities indicative of the processing history of the part. Examples include melt pool morphology, temperature, cooling rates, reflected intensity, and more. Measurable material properties are just that – aspects of the final part which are relevant to scientific or engineering applications.

The term design space will be used throughout this article to discuss the set of all additive manufacturing data which can be used with data-driven methods and machine learning algorithms. A single set of additive machine parameters, observed process phenomenon, and measured material properties can be considered as a *coordinate* in the design space. As an example, consider a list of parameters for a laser powder bed fusion process, such as [Alloy Composition, Laser Power, Scan Speed, Maximum Melt Pool Temperature, Average Cooling Rate, Final Density, Young's Modulus, Ultimate Tensile Strength]. Any part which is manufactured under a single set of conditions and is observed to have a set thermal history and material properties can be considered to be manufactured *at that point* in the design space.

Representing data in the design space property will be a major pre-processing step for machine algorithms to be amenable to AM optimization.

### B. Additive Manufacturing Data Types and Formats

Data types and sources in additive manufacturing are as widespread as any field of engineering and science. Machine learning algorithms, however, typically operate using specific mathematical representations of data. It

TABLE I. A possible design space for laser powder bed fusion additive manufacturing. Any possible combination of these parameters is a point in the design space.

| Parameter | range | step size | levels |
|---|---|---|---|
| Skin/Contour/Core laser parameters: | | | |
| Power | 100-200 W | 10 W | 10 |
| Scan speed | 500-1000 mm/s | 100 mm/s | 5 |
| Spot size | 50-100 $\mu$m | 10 $\mu$m | 5 |
| Energy density | 1-5 J/mm$^2$ | 1 J/mm$^2$ | 5 |
| Build parameters: | | | |
| Polar angle | 0-90° | 30° | 4 |
| Azimuth angle | 0-180° | 90° | 3 |
| Sieve count | 0-10 | 2 | 5 |
| Amount of recycled powder | 0-100% | 10% | 10 |
| Other parameters: | | | |
| Blade direction | 0-300 mm | 10 mm | 30 |
| Transverse direction | 0-300 mm | 10 mm | 30 |
| Hatch spacing | 0.1-0.50 mm | 0.1 nm | 5 |

is important to recognize all the different sources and formats of AM data and consider how they can be coerced into use for machine learning.

Scalar data values are some of the most common and easily obtained data in additive. A single set of manufacturing parameters, such as heat source parameters, can be represented as a list of scalars. Other scalar data can include material property measurements such as strength, hardness, density, moduli, and more. A good portion of this review covers how to model relationships between manufacturing parameters and material properties. In these cases, the machine inputs will most often be represented as a vector, such as

$$\mathbf{x} = [\text{Machine Input 1, Machine Input 2}, \ldots, \text{Machine Input } n] \tag{1}$$

out of $n$ many scalar machine inputs. Vector representations of the design space are useful in determining manufacturing conditions which will result in similar results, as well as in building regression models of the AM process.

Another common data type is *time series* data. Time series data are usually collected from models of the AM process or in situ measurement. Probably the most common time series data collected in AM is temperature histories. These data can sometimes be used as-is, or operations can be performed to extract scalar data values from the time series. Scalar statistical values such as the maximum, minimum, mean, and standard deviation of a time series signal can be equally useful in machine learning applications.

Often, experimenters and modelers have quite a few data points already collected throughout the design space. They may have many vectors $\mathbf{x}_i$, each one containing machine inputs and measured or modeled material properties. It may make the most sense to represent the data as a matrix

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & y_1 \\ x_{2,1} & x_{2,2} & \ldots & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & y_n \end{bmatrix} \tag{2}$$

where the columns of $\mathbf{X}$ represent individual machine inputs or material properties and each row is a different measurement made somewhere in the design space.

Scalar, vector, and matrix representations of AM data will be among the most common types of inputs for machine learning algorithms. If a particular algorithm discussed differs from this paradigm, it will be explicitly noted.

A final concept which is core to machine learning is *covariance*. Covariance is measured between data points, instead of being a property of a single data point. The covariance between data points encodes cross-correlated information within the design space. Ways of calculating covariance are many and varied and will be explicitly discussed where they are used. While many machine learning algorithms can operate directly on machine inputs and processing outputs, it can be equally useful to calculate covariances between data points and use those as machine learning inputs.

### C. The Assumptions Behind Machine Learning

Ultimately, AM users desire a model that relates manufacturing parameters to material properties; i.e., a process-property model. In a functional form, this model can be written as

$$f(x_1, x_2, ..., x_n) = \mathbf{y}, \tag{3}$$

where the inputs $x_i$ are $n$ different manufacturing parameters and the outputs $\mathbf{y}$ are measurable properties.

Currently, these process-property relationships are often developed by connecting separate process-structure and structure-property models of individual physical processes within AM. For example, individual process-structure models are developed to relate : heat source parameters to melt pool topologies [1] using the finite element method or similar computational techniques; melt pool topologies to solidification routes using thermal- and mass-diffusion models [2]; solidification to microstructure evolution using phase field methods [3]. Separately,

structure-property models are developed such as relating grain orientations and stress to material properties using crystal plasticity [4]. It is then by making connections between individual process-structure and structure-property models that researchers relate process parameters to material properties. This understanding-through-sequential-modeling approach has the added detriment that the errors and uncertainty of each approach compound on one another, so that the final result is less accurate, more time consuming, and more expensive than the direct build-break approach to develop process-property relationships. Machine learning is not a complete replacement for these traditional approaches, but rather a complementary modeling approach that can accelerate or even automate that process of building connections across many degrees of freedom that span the many different individual phenomenon within AM processes.

### D. Unsupervised Machine Learning

Unsupervised machine learning algorithms are used to identify similarities or draw conclusions from unlabeled data by relying on the locality hypothesis. Consider an experiment that varies three different manufacturing inputs $x_1, x_2, x_3$ and measures a single material property $y$. In matrix form, the data are expressed as:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} \\ x_{1,2} & x_{2,2} & x_{3,2} \\ \vdots & \vdots & \vdots \\ x_{1,m} & x_{2,m} & x_{3,m} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{4}$$

where $x_{i,j}$ is the $j^{th}$ measurement of the $i^{th}$ manufacturing input. A distance metric can be defined between data points in the design space. For example, data can be collected at two points $\mathbf{a} = (x_1, x_2, x_3)$ and $\mathbf{b} = (x_1 + \delta, x_2, x_3)$ and treat these quantities as vectors. Computing the $\ell_2$ norm of $\mathbf{a} - \mathbf{b}$ yields

$$||\mathbf{a} - \mathbf{b}||_2 = \delta. \tag{5}$$

The value and magnitude of $\delta$ gives an inclination about how similar $\mathbf{a}$ and $\mathbf{b}$ are. If $\delta$ is close to zero, then a researcher can say that they are similar, or even the same if $\delta$ is exactly zero. As $\delta$ becomes larger a researcher can say $\mathbf{a}$ and $\mathbf{b}$ become more dissimilar. The concept of 'similar' manufacturing conditions may be easy to assess by an experimentalist when tuning only a few parameters at a time. When taking into consideration tens or hundreds of design criteria, sometimes with correlated inputs, elucidating similar manufacturing conditions becomes difficult. This vector distance approach is a simple, yet effective first glance at similarity in a design space and is generalizable to $n$ many design criteria.

Let us say that $\delta$ is small and that $\mathbf{a}$ and $\mathbf{b}$ are similar manufacturing conditions. Now, consider a third point in the design space $\mathbf{c} = (x_1 + \delta, x_2 + \delta, x_3)$ that has not yet been measured. Since $\mathbf{c}$ was manufactured at similar conditions to $\mathbf{a}$, as measured by $||\mathbf{c} - \mathbf{a}||_2 = 2\delta$, then we may say that $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are all similar to each other. If the locality hypothesis is correct then manufacturing with conditions $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ should yield similar measurements of $y$.

At some point, a researcher will have a set of initial manufacturing inputs $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{d}$, etc., and associated property measurements that have been tested. Churning through the remainder of all possible manufacturing conditions becomes expensive and tedious quickly. Instead, researchers can use similarity metrics to determine whether or not a future test is worth running. Comparing the manufacturing inputs through vector distance gives a rough idea of the possible outcome before spending time and resources on running a test. If the intent is exploring design spaces then manufacturing at conditions *furthest away* from previously observed points may be the answer. If looking for local maxima of quality, an operator would want to manufacture at conditions *nearest to* the conditions currently known to have high quality.

Using vector distances as metrics of similarities can produce results that are analogous to creating process maps [5]. Process maps are used to divide 2 dimensional plots of manufacturing inputs into regions of quality, or regions of different material responses. The following demonstration is based on $k$-means clustering, a commonly used unsupervised machine learning clustering algorithm.

A researcher has acquired the datasets in Eqn. 4 and wants to partition $\mathbf{Y}$ into groupings of high quality parts and low quality parts. However, there are several values of $y \in \mathbf{Y}$ that lie between two extremes and the cutoff for quality is not well defined. It would be useful to use similarity metrics to find the best possible partition of quality. To begin, the data set is partitioned randomly into two groups, $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The centroids $m_1$, $m_2$ (or centers of mass, in engineering) of each grouping can be calculated as

$$m_1 = \frac{1}{|\mathbf{Y}_1|} \sum_{y_j \in \mathbf{Y}_1} y_j$$
$$m_2 = \frac{1}{|\mathbf{Y}_2|} \sum_{y_j \in \mathbf{Y}_2} y_j. \tag{6}$$

where $|\mathbf{Y}|$ is the average value of a grouping. The measurements were randomly partitioned at first; the goal

is to re-partition each set so that similar measurements (similar levels of quality) are in the same set. To do this, we can re-assign each set by

$$\begin{aligned}\mathbf{Y}_1 &= \{y_i : ||y_i - m_1||_2 \leq ||y_i - m_2||_2\} \\ \mathbf{Y}_2 &= \{y_j : ||y_j - m_2||_2 \leq ||y_j - m_1||_2\}.\end{aligned} \tag{7}$$

We can interpret the re-assignment in Eqn. 7 physically: if a measurement initially assigned to set $\mathbf{Y}_1$ is closer in distance to $\mathbf{Y}_2$ then it is *more similar* to the other set. Thus, it is re-assigned. Since the original partition was random it is likely that there are low quality parts mixed in with high quality parts - in other words, outliers exist in each partition. Measuring the similarity of each data point to the mean of the groupings reclassifies these outliers into groupings that are more reflective of their quality.

Once re-assignment is complete the centroids in Eqn. 6 can be re-calculated and updated. Then, data points are re-assigned once more based on how similar they are to the centroid of each partition. If we have partitioned the input settings $(x_1, x_2, x_3)$ along with their corresponding measurements, then we have lists of input settings which are likely to give good/bad quality parts. Further analysis can also be conducted, such as analyzing which regimes of inputs lead to good or bad quality - this is precisely what process maps represent. The difference in this case is that $n$ many manufacturing conditions can be related to a quality metric simultaneously, with little to no human inspection or intervention. Additionally, a researcher can dig further and analyze *why* groups of input settings result in given quality for a material property.

### E. Supervised Machine Learning

In a *supervised machine learning algorithm* the goal is to find a functional relationship that best approximates the underlying physical relationship $f(\mathbf{x}) = \mathbf{y}$. That is, supervised machine learning algorithms relate model inputs to labeled data. It is not necessary to make any assumptions up front about the nature of $f(\mathbf{x}) = \mathbf{y}$ to find $T$. We only need to rely on the relational hypothesis and assume the relationship exists. Functional relationships can take many forms, depending on the specific supervised ML algorithm being used. One method is to model the relationships as a vector product

$$\mathbf{X}T = \mathbf{Y}. \tag{8}$$

where $T$ is a vector of coefficients that weigh the machine inputs to approximate an entry in $\mathbf{Y}$.

A researcher usually seeks this relationship through the measurements they have observed; in this case, the measurements are stored in the matrices of Eqn. 4. A common method to find a vector representation of $T$, and a critical element in most machine learning algorithms, is through least squares regression. Least squares regression finds $T$ through a minimization problem, given by

$$\min ||\mathbf{X}T - \mathbf{Y}||_2^2. \tag{9}$$

Equation 9 can be interpreted analogously to similarity measurements for unsupervised algorithms: the closer that $\mathbf{X}T - \mathbf{Y}$ is to zero, the more similar $T$ is to $f(\mathbf{x})$.

The methods of solving equation 9 are many and varied; indeed, much of this review will focus on finding solutions to Eqn. 9 to various problems through additive manufacturing. The result is an approximation to the functional relationship $f(\mathbf{x}) = \mathbf{y}$. A new point of interest in the design space $\mathbf{x}'$ can be chosen and its associated material property $\mathbf{y}'$ can be predicted by computing

$$\mathbf{x}'T = \mathbf{y}'. \tag{10}$$

This simple example demonstrates how functional relationships can elucidate more information about design spaces. Commonly used machine learning algorithms in materials science and engineering are given in Table II. Note that the field of Machine Learning is evolving as fast as AM itself; hence Table II is by no means a comprehensive review of all machine learning algorithms, but rather is intended to provide a comparison between the form and function of some of the most widely adopted algorithms in materials science and engineering.

## III. CURRENT ICME TOOLS ARE WELL EQUIPPED TO INTEGRATE WITH AN ML FRAMEWORK

The interdisciplinary nature of AM has naturally led to a compartmentalization of scientific efforts. Many scientists choose to focus on a single problem or process because of the richness of problems in AM, even those that are narrow in scope.

The following section details how machine learning approaches can tie-in to current efforts in AM. Since the study of AM is often focused around specific problems, this section is tailored to many common areas of study within AM. Different methods of analysis and characterization for different aspects of AM all tie-in well with different machine learning algorithms. Even more so, ML can be used to automate the generation of knowledge about AM process-structure and process-property relationships.

### A. Pre-Build Design

#### 1. Alloy Design

Choice of alloy impacts the physics of AM from start to finish, starting with the optics of energy sources incident on feedstock and ending with the material properties of

TABLE II. Several of the most widely used machine learning algorithms that have been used in materials science are compared.

| Class of Algorithm | Examples | Applications | Strengths | Constraints |
|---|---|---|---|---|
| Weighted neighborhood clustering | Decision trees, k-Nearest neighbor | Regression, Classification, Clustering and similarity | These algorithms are robust against uncertainty in data sets and can provide intuitive relationships between inputs and outputs. See Ref. [6] for a primer on clustering. | They can be susceptible to classification bias toward descriptors with more data entries than others. |
| Nonlinear dimensionality reduction | t-SNE, Kernel ridge regression, Multidimensional metric scaling | Dimensionality reduction, Clustering and similarity, Input/output visualization, Descriptor analysis, Regression, Predictive modeling | These algorithms are robust against nonlinear input/output relationships and can provide intuitive projections of the material input/output space. For accessible examples, see Refs. [7, 8]. | Projections can represent unphysical, difficult to interpret relationships. Global relationships can also be lost when nonlinear dimensionality reduction results are projected onto lower-dimensional spaces. |
| Linear dimensionality reduction | Principle component analysis (PCA), Support vector regression (SVR) | Dimensionality reduction, Clustering and similarity, Input/output visualization, Descriptor analysis, Regression, Predictive modeling | This type of algorithm can produce orthogonal basis sets which reproduce the training data space. They can also provide quick and accurate regression analysis. For a primer on PCA specifically, see Ref. [9]. | The relationships studied must be linear in nature, and these algorithms are susceptible to bias when descriptors are scaled differently. |
| Search algorithms | Genetic algorithms, Evolutionary algorithms | Searching a material space to optimize on a certain condition, Lowest-energy state searches, Crystal structure prediction | Search algorithms are intuitive for material properties that can be described geometrically, such as topology optimization for weight reduction. They are efficient at searching spaces with multiple local extrema, such as finding local maxima of quality in multidimensional design spaces. | These algorithms are highly dependent upon selection and mutation criteria. For a useful application of genetic algorithms to process characterization, see Ref. [10]. |

the final part. The reflected/absorbed intensity of lasers on powder beds is determined by the powder's composition [11, 12]. The density of powders, both intra- and inter-granular density, plays a role in final part density [13]. Conduction modes in the melt are partially determined by the thermal properties of the alloy [14]. All of this is not to mention that different alloys exhibit different grain morphologies and therefore microstructures [15].

Problems in the additive process can also be linked to composition such as vaporization of constituent elements due to rapid thermal fluxes, impacting the stoichiometry of melt pools and, ultimately, quality [16]. Even traditional engineering alloys sometimes need to be altered to improve compatibility with AM. Searching for new alloys specifically for AM may also be fruitful, as unique strengthening mechanisms can arise [16, 17]. Designing alloys for AM – either altering currently used alloys or starting from scratch – requires taking into consideration the compatibility of alloys' physical properties with AM. Alloy design for AM must take into consideration general alloy properties, like melting point, to feedstock-level properties, like vaporization temperature, to bulk alloy properties, like strength. All of this information has been documented and exists in databases that can be searched. Alloy designers from AM should take advantage of these databases to search for new alloys for use in additive.

Databases exist which contain alloy properties ranging from the reflectivity of the alloy to the mechanical properties of alloys in bulk. The International Crystal Structure Database (ICSD) contains crystallographic information for millions of compositions. The Linus Pauling files contains a range of material information, from atomic properties like radius and electron valency to crystallographic level information [18]. In modern day, large databases such as AFLOWLib [19], the Materials Project [20] and more allow users to search through large databases of relevant alloy information to find one that matches a desired property. Searching through large databases of informa-

tion to find optimal compositions for manufacturing is actually one of the earliest materials informatics problems ever addressed. Methods exist to perform these searches in a fast, automated way. These methods are referred to as database mining, a data-driven materials design approach.

A study by Martin et al. used database mining to find micronucleants for Al alloys in powder bed manufacturing [14]. Part of the design process is identifying which alloy properties are important for the desired application. Heterogeneous nucleation of Al grains was the desired outcome of Martin's study. To induce such nucleation, Martin et al. searched for possible nucleants whose crystallographic lattice parameters closely matched that of Al. This way, the Al grains would have a low-energy-barrier nucleating site from which to grow heterogeneously. Martin's study employed a search algorithm to search through 4,500 different possible nucleants and identify those with the closest-matching parameters. Ultimately, Zr was found to be the best candidate.

The same process employed by Martin – identify the properties which need to be satisfied, then search for a material that is closest matching – can be extended to many other alloy properties relevant to AM. Database mining was first introduced in material science to predict stable compositions, or estimate material properties from composition. Database mining has been successfully implemented to predict stable crystal structures [21–23] and predict material properties as a function of composition [24–28]. Some specially designed search algorithms have also been designed for improved speed in automated searches [29]. Successes have been found in designing Heusler compounds using high throughput search methods [30]. Reviews of early high-throughput searches for compositions with ideal properties can be found in [31, 32]. The same search algorithms employed in these studies can be extended to the additive case.

A limiting factor in database mining is that designers are limited to properties which have been measured or calculated. We do not have information about the vast space of *possible* materials. Consider a set of alloying elements for Ti such as $\{\mathrm{Al}, \mathrm{V}, \mathrm{Zr}, \mathrm{Cr}\}$. Researchers may need to test the impact of alloying composition on the dendrite arm spacing of Ti alloys. Phase field models exist which simulate the growth of and measure dendrite arm spacing as a function of a continuum of composition. A particularly efficient combined phase field/cellular automata model was implemented by Tan et al. precisely to model dendrite arm spacing in laser manufactured alloys [2].

Modeling all possible combinations of $\{\mathrm{Ti}, \mathrm{Al}, \mathrm{V}, \mathrm{Zr}, \mathrm{Cr}\}$ is possible with coarse additions of alloying elements, but undesirable. Machine learning can aid in the process to find an optimal composition without modeling all possibilities. *Genetic alogrithms* (GA) can search the space of possible alloys to find the optimal dendrite arm spacing. Genetic algorithms have been one of the most-used data driven approaches in materials science over the past few decades [23, 29, 33–37].

The principle of genetic algorithms is to evaluate the *fitness* of a population of candidate alloys against a *fitness function*. The fitness function is a method of evaluating how well a candidate alloy meets a criteria – in this case, dendrite arm spacing.

As a thought experiment, consider using Tan's phase field/cellular automata model as a fitness function. It can be run for candidate compositions – in this case, various amounts of $\{\mathrm{Al}, \mathrm{V}, \mathrm{Zr}, \mathrm{Cr}\}$ alloyed into Ti – and used to evaluate the dendrite arm spacing. Once a fitness function has been identified, the next step in a genetic algorithm is to represent candidate alloys as a *gene*.

We can represent a gene as

$$\text{Alloy} = [\chi_1,\ \chi_2,\ \ldots,\ \chi_n]$$

where $\chi_1$ is the species and weight percent of the first element (titanium, in this example), $\chi_2$ is the species and weight percent of the second element, up to $n$ elements. For example, Ti-6Al-4V would be represented as

$$[0.9\ \mathrm{Ti},\ 0.06\ \mathrm{Al},\ 0.04\ \mathrm{V}\ ]$$

The goal is to find the alloy with optimal dendrite arm spacing. First, a population of candidate genes needs to be generated, either randomly or by design. Two examples from a starting population may be

$$\begin{aligned}\text{Alloy 1} &= [0.9\ \mathrm{Ti},\ 0.05\ \mathrm{Al},\ 0.05\ \mathrm{V}\ ]\\ \text{Alloy 2} &= [0.9\ \mathrm{Ti},\ \ 0.1\ \mathrm{Zr}]\end{aligned}$$

The dendrite arm spacing can be estimated using Tan's model. It is not guaranteed that the optimal composition is contained in this starting population.

Genetic algorithms select genes out of the current population – called the parent generation – to proceed to another generation of model assessment – called the child generation. Selection consists of keeping the best performing compositions, say the top 10%, and discarding the rest. Genetic algorithms find optimal locations in the design space by relying on the similarity hypothesis. If one alloy is in the top 10% of genes then it is possible that a similar alloy will also be high performing – it may be even perform better. Once selection is done, the next step is to search the space near the best performing alloys from the parent generation.

Genetic algorithms generate similar compositions from those selected in the parent generation by making alterations to genes. One operation is *mutation*, whereby entires of the genes are changed. For example, we could mutate alloy 1 by changing the composition:

| | |
|---|---|
| **Parent Generation:** | Alloy 1 = [0.9 Ti, 0.05 Al, 0.05 V ] |
| **Child Generation:** | Alloy 1 = [0.9 Ti, 0.02 Al, 0.08 V ] |

where in the child generation the amount of V was increased, while the amount of Al was decreased. Another operation which may be performed is *crossover* where entries of genes are added or interchanged. For example, one crossover operation may look like

| | |
|---|---|
| **Parent Generation:** | Alloy 1 = [0.9 Ti, 0.05 Al, 0.05 V ] |
| | Alloy 2 = [0.9 Ti, 0.1 Zr] |
| **Child Generation:** | Alloy 1 = [0.9 Ti, 0.05 Al, 0.05 Zr ] |
| | Alloy 2 = [0.9 Ti, 0.1 V] |

where in the second generation V and Zr have been interchanged.

Selection, mutation, and crossover followed by model assessment and then further selection, mutation, and crossover continues until the design criteria is met. The intuition behind genetic algorithms is that eventually the selection process is narrowed down to alloys within a given region such that further mutation and crossover do not produce new genes. Eventually, all the 'fittest' genes as determined by the model will converge to be approximately the same composition.

Genetic algorithms have been applied to alloy design for low and high temperature structural materials [24, 38], ultra high strength steels [39], specific electronic band gaps [40], minimum defect structures [41], exploring stable ternary or higher alloys alloys [35, 42], and more. For a review on the application of GA's to alloy design through the early 2000s see Ref. [43].

Other machine learning algorithms have also been applied for classification and optimization of alloy compositions. Anijdan used a combined genetic algorithm–neural network method to find Al-Si compositions of minimum porosity [41]. Decisions trees have been implemented for a number of different alloy optimization, such as predicting ferromagnetism [44] and the stability of Heusler compounds [45].

Of course, some compositions definitely *won't* be compatible with the additive process. It would be useful to identify these alloys up front or as quickly as possible. Additive can also be improved through by machine learning algorithms which suggest the best materials or properties to test. A wide range of machine learning algorithms can be implemented to guide the entire experimental design process so that an optimized property is found as quickly as possible.

### 2. *Design of Experiments*

Parametric analysis, broadly defined, is an experimental method of mapping independent variables to their corresponding dependent parameters. Process-property relationships are typically studied through parametric analysis. Machine learning aids in investigations of AM by reducing the amount of experiments needed to characterize process-property relationships. Machine learning approaches like sequential learning model relationships in parametric studies to discover regions of the parameter space which will produce the most information about process-property relationships.

In additive manufacturing independent parameters such as laser energy, speed, build direction, composition, layer height, and more are varied to study their impact on material properties. Examples include relating build geometry to microstructure or surface roughness [46, 47] or temperature history to microstructure [48, 49], or substrate temperature to residual stress development [16, 50], or even entire manufacturing processes to microstructure [51]. One of the most common types of parametric analysis is relating heat source parameters to all aspects of AM, such as part temperature history [52, 53], microstructure [54, 55], mechanical properties [56, 57], residual stresses [58, 59], and more.

Both engineering and scientific investigations of AM utilize parametric analysis. In the sciences, parametric analysis proceeds until a theory or model can be presented for a process-property relationship. In engineering, parametric analysis continues until an optimality criterion is met, such as maximum strength or minimum porosity. Both disciplines vary independent parameters and measure dependent responses; doing so provides information about the underlying phenomenon.

*Information* is any observation of process-property relationships. For example, observing that a set of laser parameters results in an equiaxed microstructure can be considered information because the researcher has gained an idea of the properties to expect from set processing conditions. Therefore, *information gain* is any experiment which reveals a previously unobserved process-property relationship. Rigorous mathematical definitions of information and information gain have been defined, typically referencing back to Shannon's original formulation of information theory [60].

Traditional design of experiments maximize information gain by dividing up parameter spaces to maximize distance between like experimental conditions. Machine learning driven design of experiments makes suggestions for future experiments based on the results of past experiments. It does not make any assumptions up front about correlations between design parameters. Rather, machine learning algorithms model relationships between inputs and outputs and suggest the experiment which is statistically most likely to result in information gain. Design of experiments with machine learning algorithms can be adopted by augmenting traditional design of experiments with a statistical model.

The first step is to define the space of parameters which may impact properties, as is done in traditional design of experiments. The design space is all possible combinations of process parameters. As more parameters and finer step sizes are added, the size of the design space grows. Once the design space has been defined, the next step is to generate an initial dataset. The process-property relationships revealed in these initial tests will be the basis of an initial statistical model.

After an initial dataset is generated, the researchers need to define a *response function* which interprets the relationship between parameters and material properties. One example is a regression model of the process parameters and material characteristics. Identifying tests of importance begins by identifying all $n$ parameters in the design space that may or may not impact a final manufacturing result $y$. First, a subset of the $n$ many parameters $m_1$ is chosen such that at least one parameter in $\mathbf{x}$ is left out. A simple regression function called a decision tree is trained

$$\hat{f}_1(\mathbf{x}_{m_1}) = y \tag{11}$$

on the subset of features $m_1$. Next, a new random sampling of the $n$ parameters $m_2$ is chosen, and a second regression problem is solved $\hat{f}_2(\mathbf{x}_{m_2}) = y$. The random forest is a linear combination of all of these individual regression functions

$$\hat{f}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\mathbf{x}_{m_b}) \tag{12}$$

out of $B$ many parameter samplings. A standard deviation in the prediction output is typically computed as

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} \hat{f}_b(\mathbf{x}_{m_b}) - \hat{f}(\mathbf{x})}{B-1}} \tag{13}$$

Some of the features in $\mathbf{x}$ may be irrelevant for studying process-property relationships and some may be very important to model. Random forests test the importance of an AM process parameter in $\mathbf{x}$ by re-training every individual regression function $\hat{f}_b(\mathbf{x}_{m_b})$ with that machine input included. Then, the standard deviation is re-computed. If the inclusion of a parameter in every regression tree significantly changes $\sigma$ then it plays a large role in prediction of $y$. Therefore, the feature should be included and tested in simulations.

Ling et al. performed design of experiments to optimize the process of designing high fatigue strength steels. [61]. In Ling's case, a random forest model was fit to design criteria from a database of steel compositions, processing routes, and fatigue strength. The goal was to find the composition and processing route combination which had the highest fatigue strength in the dataset using as few experiments as possible. A benefit of random forest models is that they provide uncertainty estimates on predictions. This means that regions of the design space with high uncertainty can be identified.

Random forests are easy to use because they do not require tuning many hyper-parameters. Furthermore, random forests are very good at sifting through many irrelevant features to find the features that actually matter. Random forest training is also computationally inexpensive and easily parallelized. Random forests provide two advantages that are particularly important in the context of materials science: the ability to efficiently calculate uncertainty estimates and the added interpretability of feature importance metrics. Based on their ensemble nature, it is possible to generate uncertainty estimates for random forest predictions using jackknife-based procedures [62–64]. They automatically identify the most important features in a given model based on how often a feature is used in splitting criteria and what the aggregated information gain is over those splits. These feature importance metrics can provide insights into how the model is making its predictions. Random forests have been applied successfully to a range of applications in materials science. They have been used to discover new Heusler compounds [45] and new thermoelectric materials [65]. They have also been used to model material properties such as thermal conductivity in half-Heusler semiconductors [66] and to break down fields for dielectrics [67].

Ling's machine learning-assisted design of experiments proceeded by suggesting experiments with a high uncertainty in their result as modeled by the random forest regression. The intuition is that predictions by the response function which have low uncertainty have enough data to characterize the process-property relationships in that region of the design space. Therefore, researchers can have high confidence in the material properties they will achieve if parts are printed at those conditions. Thus, the process-property relationship likely has enough information to pick an optimal condition or investigate further. Regions of the design space with high uncertainty do not have enough information therefore further experiments are required. Thus, machine learning algorithms suggest these regions of the design space for further experiments. When only varying a few parameters at a time, regions of the design space which need characterization can be easily identified. When varying tens of parameters in additive manufacturing these regions of the design space are not apparent. Furthermore, correlated inputs can be masked by the complexity of the process-property relationship. Statistical models can spot these correlated regions of the design space.

A regression model, such as that used by Ling, is only one way of assessing process-property relationships. A review article detailing many optimization algorithms for design of experiments can be found in Shan et al. [68]. Adoption of machine-learning assisted design of experiments algorithms can rapidly increase the rate at which the relationship between AM process parameters and material properties are understood.

### *3. Topology Optimization*

Alloy design and experimental design focus around combinatorial screening of inputs to either search for *new* properties or optimize on current properties. These optimizations reduce manufacturing cost, monetary or otherwise, and maximize performance capability. The same optimization can be applied to mechanical properties of parts. For structural materials, the goal is to optimize

load bearing capacity or lifetime while minimizing the amount of material used. For aerospace, the goal is to minimize weight. Unique manufacturing geometries was one of the first intended applications of AM. Topology optimization (TO) focuses around exactly this task – finding optimized topological structures for a given mechanical application.

A filter is a mathematical operation which reveals information about a region of pixels/voxels in a mesh. Filters are most often represented as a product of a filter matrix with a matrix of mesh pixel values. Topology optimization proceeds by generating a CAD model of an AM part and modeling its performance, such as testing performance under mechanical load through an FEA simulation. Filters are applied to the CAD mesh which selectively removes material from the part. Then, the mechanical performance of the new part is modeled, followed by further material removal. This process proceeds until either a minimum weight/volume condition is met or the mechanical performance of the part is degraded.

In additive, topology optimization serves an additional purpose: TO algorithms can find un-printable regions of a part. Unsupported structures, low angle slopes, and certain part orientations during building are prohibited in AM because they will cause part deformation. An unsupported slope at acute angles can lead to part deformation and warpage [69]. Sacrificial support structures also need to be considered during topological optimization, along with the number of free-hanging features and the orientation of the part during manufacture. Langelaar et al. developed an AM-specific TO algorithm which searches for regions of parts that have too little support for manufacture [70, 71]. Other additive specific algorithms have been designed for optimizing density of parts [72]. These algorithms augment the AM process, both by taking advantage of the ability to optimize unique geometries, and also by identifying regions of parts which are incompatible with AM.

## B. Process Design

Integrated computational materials engineering, as the name may imply, is primarily focused around *computational* engineering of materials. The design space in AM makes choosing useful experiments difficult; the wide range of physics of AM makes full-scale, full-physics computational modeling a burden. This section focuses around applying machine learning algorithms to aid in computational studies and design of additive manufacturing.

### 1. Model Complexity and Dimensionality Reduction

A bottleneck in ICME approaches is the number of different models required to simulate all physics during additive manufacturing. The spatiotemporal scales that *could* be taken into consideration for AM process modeling span microns to meters. ICME of AM process modeling incorporates phase field, cellular automata, finite element, direct element methods, and more. At the smallest scale the feedstock, heat source, and melt pool dynamics have been modeled by finite element methods [1, 73, 74] or finite volume methods [75]. Microstructure growth from the melt pool is often the next phenomenon to be studied and has been modeled through phase field [3, 76–78], cellular automata [2], or finite element methods [49]. Thermal histories of entire parts or sections can be modeled next, typically by finite element methods. Thermal history models look at heat transfer through the part [79], residual stress build up during manufacturing [4, 80], and thermal history such as cooling rate and temperature gradient [53, 81]. Full ICME approaches involve modeling the AM process at all these steps to determine the entire processing history for a part. Martukanitz et al. published a full ICME investigation of AM [82]. A review of ICME approaches across spatiotemporal scales can be found at [83] and a review of the physics of AM modeling can be found at [84]. A review of finite element methods specifically for AM can be found at [85].

For many engineering studies of AM not all of these physics *need* to be modeled or included. The complexity of AM, however, obscures which physics are relevant to a given final property. When considering the design space of additive, listing out the possible parameters and properties to model illustrates the complexity of modeling. A data point in the AM process could be written as a vector

$$\begin{aligned}\mathbf{x} = (&\text{Laser Power}, \text{Laser Speed}, \text{Laser Spot Size}, \dots \\ &\text{Energy Density}, \text{Alloy}, \text{Build Time}, \dots \\ &\text{Build Orientation}, \text{Hardness}, \text{Ultimate Tensile Strength}, \dots \\ &\text{Toughness}, \text{Porosity}, \text{Surface Roughness}, \dots ).\end{aligned} \tag{14}$$

By some estimates, the vector $\mathbf{x}$ could be 160 variables long if only accounting for tunable machine parameters. Complexity in ICME of AM only grows when considering cross-correlated information across different models. The result of one simulation $y$ could be an input $\mathbf{x}$ of another simulation. A thermomechanical model may predict temperature gradient in deposited layers, while a phase field model may use thermal gradient to calculate solidification velocity. This problem motivates using computational models which balance accuracy with computational complexity.

Dimensionality reduction algorithms identify which parameters are relevant to model in an ICME approach and which are not, informing future ICME investigations of faster simulation routes to achieve the same result. Materials science has long had a need for dimensionally reduced, computationally accurate models. Some of the first applications of machine learning in materials science was for dimensionality reduction [22, 86–88]. The dimensionality reduction techniques covered in this review

can be broadly classified into three mathematical frameworks: statistically driven methods, similarity analysis, and matrix factorization.

Statistically driven approaches can be employed to determine the parameters in $\mathbf{x}$ that strongly impact AM model outputs, such as thermal history or materials property. A commonly used, statistically driven dimensionality reduction method is through random forest networks, as is done in design of experiments. The difference, however, is how uncertainty in the random forest is employed. In design of experiments, inputs or outputs with high associated uncertainty in the random forest model may be chosen for future tests. In dimensionality reduction, inputs which have high uncertainty are indicative of irrelevant features to model.

Kamath utilized random forests to screen out irrelevant modeling parameters for predicting maximum density of additively manufactured parts [89]. Kamath started with an experimental dataset of manufacturing parameters and multiple modeling methods. An Eagar-Tsai simulation of a Gaussian laser beam on a powder bed was used to model thermal conduction during manufacture. The model originally began with four inputs (laser power, speed, beam size, and powder absorptivity) and a design space of 462 possible input combinations. Kamath utilized random forests to determine which input was most important for achieving fully dense parts. If simulations are expensive to run then 462 different simulations may be out of the question. Identifying which parameters do not impact the final result reduces the size of input combinations, therefore reducing the number of computations to be performed. As a fictitious example, if the angle between the laser and part surface does not impact the result of a simulation, then simulations do not need to be run which vary the angle. This is valuable, time-saving information in an ICME study.

Kamath identified that laser speed and power were the most important inputs out of the four to determine melt pool depth and shape. Thus, model predictions made varying only these parameters can be relied upon for their results. After determining the most important inputs, the same regression tree was applied in order to find optimized manufacturing conditions for fully dense parts, using the same approach. Instead of identifying which features impacted the model standard deviation, the machine settings which maximized $y$ were found.

Dimensionality reduction through similarity analysis allows visual interpretations of distributions in high dimensional space. Understanding an $n$-dimensional distribution is difficult. Process maps are commonly employed in AM to ascertain the distribution of material properties with processing parameters [5]. Process maps are 2D plots which chart the possible values of machine inputs and identify regions of the design space with similar properties. A commonly employed process map in AM is the divide between grain morphology as a function of solidification velocity $R$ and temperature gradient $G$ [90]. Plotting the distribution of $n$ many process variables would equate to $\binom{n}{2}$ plots. Similarity analysis can be used to express distributions measured in an $n$ dimensional space in a human interpretable way without relying on multiple 2D process maps.

The algorithms discussed thus far have used AM data 'as is' by using machine inputs and materials properties to train the model. It can be equally useful to train models on *correlations* between machine inputs, instead of inputs themselves. This approach combines similarity analysis with regression. Covariance between AM parameters are measured by similarity and these similarity scores are the data for a regression model.

$t$-distributed Stochastic Neighborhood Embedding (tSNE) is a dimensionality reduction technique which measures distances in a high dimensional space and then projects data points onto a two dimensional plot. The similarity of all data points in the design space with each other can be used to fit a distribution of similarities. The tSNE algorithm begins by fitting a probability distribution to all $\mathbf{x}$'s contained in a dataset. Relationships in $n$ dimensional space are assessed through a *kernel function* $\kappa(\mathbf{x}, \mathbf{x}')$ which measures similarity between points in the design space. A commonly employed kernel is the Gaussian kernel

$$\kappa(\mathbf{x}, \mathbf{x}') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\mathbf{x} - \mathbf{x}'}{2\sigma^2}\right] \tag{15}$$

where $\sigma$ is a user-specified or fit standard deviation in the distribution of points in the design space. This kernel function assesses distance in the $n$ dimensional space and assigns a similarity value between $[0, 1]$ depending on how similar the values are.

After the $n$ dimensional dataset is fit, then a 2 dimensional coordinate $\mathbf{x}^*$ is assigned to each $\mathbf{x}$. The reason for choosing a 2 dimensional coordinate is so that the final result can be visualized on a 2D plot. The tSNE algorithm works by fitting a probability distribution to the $n$ dimensional data set first, then assigning values to each $\mathbf{x}^*$ such that they have the same probability as their associated high-dimensional $\mathbf{x}$. Once the probability distributions have been assigned, the $\mathbf{x}^*$ values can be visualized on a 2D plot to investigate for trends.

The benefit of tSNE is that points which are close together in the $n$ dimensional space appear close together on the 2 dimensional plot. This gives AM modelers an idea of how machine inputs and material behavior are distributed in the $n$ dimensional space through a 2 dimensional visualization. This dimensionality reduction technique guides modelers as to what simulation conditions will reveal similar results. Furthermore, these 2D projections provide clusters of inputs and outputs which are similar in the high dimensional space. Analysis of these clusters can inform modelers as to correlations within their inputs and outputs, identifying which parameters can be screened out to save complexity.

A final dimensionality reduction technique requires expressing model data in a matrix and performing matrix factorization. For a dataset of model inputs and results,

a matrix can be formed $\mathbf{X}$ whose rows are the machine inputs, calculated properties, and simulation conditions contained in the vector $\mathbf{x}_i$. Matrix factorization techniques re-represents correlations in large datasets in a simplified way. The matrix $\mathbf{X}^T\mathbf{X}$ is a measure of covariance within $\mathbf{X}$, different than Eqn. 15 but equally valid. The matrix $\mathbf{X}^T\mathbf{X}$ can be very large due to the design space of additive manufacturing. One type of matrix factorization, called Principal Component Analysis (PCA) re-represents the data matrix $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}^T \tag{16}$$

where the rows of $\Sigma$ are the eigenvectors of $\mathbf{X}^T\mathbf{X}$ and the diagonal entries of $\mathbf{U}$ are the square root of the eigenvalues $\sqrt{\lambda}$. PCA operates such that the first eigenvector in $\Sigma$ indicates the most heavily correlated inputs of $\mathbf{X}$. The length of vectors in $\Sigma$ is $p$ many variables long. In cases where the original design space size $n$ is large, the size of $p$ is often much less than $n$, reducing the dimension of the problem. Regression can be performed on one, a few, or many of the eigenvectors in $\Sigma$ to predict new model results using considerably less information than that contained in $\mathbf{X}$. Some studies have gone as far as to understand which correlations are being represented by the eigenvectors in $\Sigma$, revealing inputs or phenomenon which can be ignored during modeling. Such analysis is possible in AM, though requires further processing steps past matrix factorization. Materials science studies have utilized PCA previously to re-represent large datasets in simpler forms, such as predicting the formation energies of crystal structures from a lower dimensional space [91]. A review of applications of PCA in materials science can be found at [92].

In additive manufacturing, PCA can serve as a dimensionally reducing pre-processing technique. For a large data set with many machine parameters the design space can be reduced from vectors of $n$ length to $p$ length. The new vectors can then be used as regression model inputs, further reducing the computational complexity of AM.

#### 2. *Surrogate Modeling*

Dimensionally reduced models are useful for engineering applications where few properties are being studied at a time, or when computational burden hinders development time considerably. Full-physics modeling is quite necessary to understand how physics at different scales interact to impact the AM process. Full physics models, however, can be expensive and time consuming to run. Phenomenon which are difficult to study experimentally, such as flow within the melt pool, are best studied through modeling approaches. If a model's computational expense is *too* high then performing simulations at all relevant manufacturing conditions can be impossible. Machine learning algorithms can use the results of previously run high fidelity simulations to fill in the gaps and reduce development time.

A *surrogate model* is a mathematical approximation that predicts the results of AM simulations without actually performing the computation. Surrogate models preclude the need for running computationally expensive simulations for every possible manufacturing condition. The results of previous data-intensive simulations can be used in regression models to predict the results a simulation *would have* given, if it had actually been performed. Surrogate models can be as simple as linear regression between simulation inputs and results, but are often more complex. The accuracy of a surrogate model is dependent upon how many previous simulations have been run and at how many different points in the design space.

Tapia et al. built a surrogate model for laser powder bed fusion of 316L stainless steel. They were concerned with predicting the melt pool depth of single-track prints solely from the laser power, velocity, and spot size [93]. The dataset used to build the surrogate was computationally derived, based on previous simulation methods used by the same research team [94]. In particular, they used the results from a computationally expensive but high-accuracy melt pool flow model of Khairallah et al. [1]. They ran powder bed simulations at various laser powers, velocities, and spot sizes, and the model told them the depth of the melt pool, among other information. The datasets provided enough information for a surrogate model to be trained to predict simulation results.

To build this model, Tapia et al. used a machine learning model known as a Gaussian process model (GPM). A common model assumption in Gaussian process modeling is

$$z(\mathbf{x}) = y(\mathbf{x}) + \epsilon(\mathbf{x}) \tag{17}$$

where $y(\mathbf{x})$ is the approximation (surrogate) of the simulation process, $\epsilon(\mathbf{x})$ is a stochastic, randomly distributed noise in measurement, and $z(\mathbf{x})$ is the value given by a simulation. GPMs make the assumption that all finite joint distributions of $y(\mathbf{x})$ are distributed in a multivariate normal manner, also known as a Gaussian distribution. For AM modeling this assumption holds if the model data used for a surrogate is stochastic in nature. The primary goal in GPMs is to find model parameters for the mean process $y(\mathbf{x})$ and a covariance function $C(\mathbf{x}, \mathbf{x}')$, which is a function of similar form to Eqn. **??**. Fitting a Gaussian process model often begins with assuming a model function for covariance, fitting the model parameters to the observed values $z(\mathbf{x})$, then using those model parameters to predict simulation results $y(\mathbf{x})$ at other locations in the design space.

Starting with the covariance of their inputs, Tapia used Bayesian statistics to develop a probabilistic model that predicted melt pool depth from simulation inputs. They were able to successfully predict the outcomes of both high-fidelity simulations and experimental measurements solely by analyzing trends in previously obtained results. In particular, they were able to accurately predict the melt pool depth at a value that had never been observed

before, either computationally or experimentally. For future investigations, predictions by the surrogate model can be relied up on instead of running a simulation or experiment.

Gaussian process models have benefits beyond their surrogate modeling capabilities. GPM provide robust uncertainty metrics on the predictions they make. Uncertainty in prediction is important in materials informatics because it aids in discriminating against poor prediction accuracy in machine learning. Some machine learning models do not have straightforward ways of assessing model error [95].

Another benefit of GPM is that it aids in inverse design and design space visualization. GPMs can explicitly identify regions of the design space which will maximize or minimize a value. In the case of Tapia et al. response surfaces were created from the GPM which visualized the depth of melt pools as a function of laser power and speed. Doing so allows engineers to identify regions of the design space which provide specific material responses, an important tool in optimization for additive.

Another approach to full scale models in AM is building high-fidelity models from an ensemble of low-fidelity models. Current integrated computational models link phenomena across spatiotemporal scales by running many single-physics models and passing the results from model to model, then comparing with experimental results. An example is the model of Martukanitz et al. that considered the thermal, mechanical, and material response of Ti-6Al-4V alloys manufactured with powder bed processes [82]. Martukanitz's model uses the finite element method to model the thermal and mechanical response based on classical continuum heat transfer equations. At the same time, the Calculation of Phase Diagrams (CALPHAD) method modeled the thermodynamics and mass transfer for each chemical species, while Diffusion-Controlled phase Transformations (DICTRA) is used to model phase formation. As noted by Martukanitz, this approach becomes computationally infeasible as the number of deposition layers increases.

Even a *single* modeling method – just FEA, phase field, CALPHAD, etc – becomes computationally expensive for thousands of layers of material deposition. Based on this scaling, accurately modeling the physics of a full build seems impossible, at least without a major leap in computing power. A machine learning method known as *committee voting* or *ensemble modeling* may provide a workaround. These methods rely on sampling many individual models to predict the behavior of a larger class of physics.

In traditional ensemble modeling approaches, many different regression models are trained to model a relationship $y = f(\mathbf{x})$, as in regression trees. For a given input $\mathbf{x_1}$ all regression models are assessed and various outputs predicted, $f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_n(\mathbf{x})$, etc. In AM it may make more sense to use data from single-scale AM models as the base models and train a machine learning based regression model from those. These data sets for additive may be the inputs and solutions of a solidification or heat transfer model, or experimentally obtained relationships.

Ensemble modeling can be used in AM to solve multi-objective optimization problems, such as optimizing the heat transfer through a build and the grain growth simultaneously. Doing so would require training an ensemble model of simulations. The simulations to include could be thermomechanical models of heat transfer, phase field models of grain growth, finite element models of laser absorption, and so forth. The final ensemble model would take the form

$$\mathbf{y}(\mathbf{x}) = \sum_{i=1}^{N} w_i f_i(\mathbf{x}) \tag{18}$$

where $\mathbf{y}$ is a response vector of properties to optimize, $f_i(\mathbf{x})$ is a single simulation of the process out of $N$, and $w_i$ are weights associated with each simulation. The vector $\mathbf{y}$ would contain many parameters such as temperature $T(t)$, temperature gradient $\nabla T$, solidification velocity $\nu(t)$, and many more phenomenon to monitor from simulations.

In the AM case, where information from models may overlap, this type of approach can screen out noise from simulations or experiments to gauge a more fundamental relationship. The uncertainties associated with a model can be weighted by the experimentally observed data points. Furthermore, predictions can be made across many different physical models, resulting in a predictive method for holistic analysis of the final properties. This method does not provide a picture of how the physics in each simulation interact or disagree. However, it can be used to simultaneously optimize the results of many different simulations and point toward desirable manufacturing conditions. A review of multiobjective optimization functions can be found at [96].

Meredig et al. applied this method to the prediction of ABC ternary compounds, where A, B, and C each represent an element. [97]. The space of all combinations of A, B, and C elements is large and would require many, many simulations. Instead of running high-throughput DFT for all possibilities they used lower order simulations. Meredig ran DFT calculations for AB, AC, and BC compounds to generate a database of information. Then, regression trees were trained from the inputs of each binary alloy simulation to predict the formation energy. Finally, an ensemble model was trained

$$E(\mathrm{ABC}) = w_{\mathrm{AB}} f(\mathbf{x}_{\mathrm{AB}}) + w_{\mathrm{AC}} f(\mathbf{x}_{\mathrm{AC}}) + w_{\mathrm{BC}} f(\mathbf{x}_{\mathrm{BC}}) \tag{19}$$

where $E()$ is the formation energy, $w_{\mathrm{AB}}$ is the weight for a regression model of AB alloys, $f(\mathbf{x}_{\mathrm{AB}})$ is the regression tree for all AB alloys, and so on. In the end, Meredig's model was able to predict $4,500$ new, stable ternary materials.

Machine learning is not only limited to ex situ experimental investigations or modeling approaches. Machine

learning has also made major advances in signal processing and feedback. Many of the same ideas which apply to experimental and modeling studies can also be aids for in situ analysis and feedback of the manufacturing process.

### C. Process Monitoring and Characterization

Monitoring of AM processes produces data equally numerous to, if not in excess of, the data produced by parametric analysis and modeling. The numerousness of time series data in AM warrants usage of quick, efficient, and robust signal processing methods for process monitoring, feedback, and control. These signal processing algorithms are closely related to machine learning and may serves as pre-processing techniques for using data in other machine learning applications like clustering and regression.

Machine learning algorithms can serve three problems in signal processing of AM:

- real time analysis, event detection, and feedback mechanisms during manufacture;
- fast analysis and quantification of image data

#### 1. *In Situ Process Monitoring and Feedback*

Computer vision is a class of image recognition algorithms that have been developed for automated feature identification in signals. Intelligent computer vision utilizes machine learning algorithms to identify objects and features in images and time-series data.

Computer vision can be employed in additive manufacturing to monitor the printing process, such as characterizing temperature profiles, identifying abnormal melt pool morphologies, and automatically detecting defect formation. Doing so will require methods for in situ process monitoring and data collection. Thus far, in situ control in AM has been consistently ranked as one of the most-needed technologies for advancing the technology [98–100]. The combination of rapid solidification and the small length scales of AM solidification can make traditional process monitoring approaches difficult. Machine learning can fill in gaps where human-specified process monitoring models are insufficient.

Process monitoring involves acquisition of realtime signals which can reveal information about manufacturing. McKeown et al. used dynamic transmission electron microscopy to measure solidification rates in powder bed AM [101]. Bertoli et al. also characterized cooling rates using high speed imaging [102]. Raplee et al. used thermography to monitor the solidification and cooling rates of electron beam powder bed fusion, then related the temperature profiles to defect and microstructural characteristics [103]. Distortion of parts due to thermal cycling was investigated by Denlinger et al. by means of thermocouples in contact with the build substrate [59]. All of these methods are amenable to aid by computer vision and signal processing.

The type of data being collected in situ is most often in the form of time series data or images. Signal processing of either typically involves identifying signals which deviate from a mean, desired signal. Examples include a spike in temperature from a thermocouple or a sharp change in intensity in an image, both of which may indicate a deviation from the desired processing conditions. Image processing tools known as *filters* can be used to identify signals of interest in AM data. From a mathematical perspective, filters are implemented for signal/image processing in the same way as for topology optimization.

A filter is typically implemented as mathematical operation on a segment of time series data or an area of pixels in an image. The simplest filters identify characteristics of the pixel intensity in that region and highlight them. An example for an image would be to identify regions of sharp change in intensity, which often correspond to boundaries between objects. An example would be the change in intensity between a liquid melt pool and the powder bed. Signals of interest or regions of interest in an image identified by filters are known as *features*.

The use of filters alone does not constitute machine learning, but filters can be used in regression algorithms to learn signal features which are indicative of behavior in the manufacturing process. Defect detection is an apt use of filters in a machine learning approach. Repeated defect-free printing involves following the same processing parameters every manufacture. Regression algorithms can predict when a defect appears by monitoring when features indicative of defects are found by filters.

A machine learning method called *template matching* can be utilized for automatic identification of signal features. Template matching involves comparison of a database of pre-identified signal features with the features measured in situ. The scale-invariant feature transform (SIFT) [104] and a variant of it, speeded-up robust features (SURF) [105] are both feature identification algorithms which can be used for template matching. A specific type of template matching is the 'bag of visual words' or 'dictionary' method. A collection (dictionary) of typical features from the AM process can be built based on features obtained from filters. The features measured in situ are compared with dictionary entries. If an in situ feature matches a defect-indicative feature from the dictionary, then it is likely a defect has formed in manufacturing process.

Template matching is carried out through similarity of analysis of in situ features and dictionary features. Outputs of filters can also be used as inputs to regression algorithms to predict the probability of a feature being present in an image. For example, an algorithm may be trained which convolves three different filters with a signal acquired in situ. Each filter likely highlights a different aspect of the signal. The goal of the algorithm

is to predict whether or not a defect is present in pixel location $P(x, y)$. A regression algorithm can be trained

$$P(x, y) = w_1 I_1(x, y) + w_2 I_2(x, y) + w_3 I_3(x, y) \quad (20)$$

where the probability of a defect being present is dependent upon the results of filter operations on regions of the image. This type of algorithm is the basis for a *neural network* where filters are used as regression functions for feature identification in images. Recently, advances in computer vision has led to the advent of *convolutional neural networks* which have high prediction accuracy for feature identification in images.

Convolutional neural networks not only adjust the weights $w_i$ of regression algorithms for feature identification, but also adjust the filters themselves. That is, convolutional neural networks 'learn' the best filters for elucidating features which are indicative of a desired feature. CNNs use large databases of labeled information to learn features in the image that indicate whether or not certain human-identified objects are present. Convolutional neural networks can identify multiple features or defects in an image simultaneously. A downside to CNNs is that they require **very** large datasets (thousands of images, at the least) of labeled data to be successful. Neural networks have already been implemented for in situ AM analysis. Yuan et al [106] were able to successfully monitor melt track width, standard deviation, and continuity of tracks in situ during laser powder bed manufacturing. Scime and Beuth trained a convolutional neural network to identify six different types of defect which are typical of laser powder bed fusion, with fairly good prediction accuracy [107]. These studies encompass only a few of the possibilities of in situ process monitoring for AM.

### 2. *Featurization of Qualitative Image Data*

The use of images in studying additive manufacturing is widespread, common throughout all aspects of the manufacturing process, and provides key information about material properties and processing. As with all aspects of AM, the sheer size of image data to be analyzed is profound due to the large design space of AM. The types of information taken from images includes grain characteristics, like size, orientation, and phase, and defect characteristics, like pore size or crack length. When characterizing all of these features for all possible processing conditions and alloys the size of the problem grows quickly.

Computer vision algorithms have been tested for automation of materials science image classification and analysis. Using these algorithms can speed up the experimental characterization process of AM. Furthermore, computer vision techniques can quantify information which may have otherwise only been used qualitatively or measured by approximation.

It is worthwhile to mention up front that these algorithms have been *tested* on microstructure and, in some cases, additive-specific images. There are few algorithms that can process AM microstructure data 'out-of-the-box.' Rather, these algorithms will need to be tailored in order to quantify AM images specifically. However, the algorithms discussed here have been proven on non-AM microstructure datasets, thus they should be extensible to AM datasets. The computer vision approaches which work for microstructure data are often the same approaches discussed in the previous section for in situ monitoring.

One AM-related application of image characterization is measuring particle size distributions in AM powder feedstock. DeCost and Holm used SIFT with a dictionary classifier, as in template matching, to measure the particle size distribution for a dataset of synthetic powder particles [108]. Particle size distribution plays in several steps across the additive process including energy absorption and part metrology [11, 109, 110]. DeCost created datasets with six different particle size distributions. Image features were identified and classified using $k$-means clustering on the features found by SIFT. Then, a classification algorithm known as a support vector machine (SVM) was trained to classify image features into particle sizes. DeCost was able to achieve 89% overall classification accuracy in measuring particle size distribution this way. DeCost et al. later improved upon this powder classification method and were able to achieve higher classification accuracies for real powder images [111].

DeCost et al. have also made strides in identifying and quantifying information from metallographs [61, 112, 113]. A good portion of quality control in materials science as a whole, not just AM, involves classifying materials based on metallographs or micrographs of microstructure. Work is being done across materials science to apply machine learning based computer vision to classifying and quantifying information in these microstructural images. Doing so will speed up the process of materials characterization and qualification, while also providing methods of quantifying information which otherwise would have stayed in a qualitative form. Examples include classification of grain stuctures, measurements of grain size, pore size calculations, and more.

Chowdhury et al. took a more expansive approach to performing feature identification in microstructures. In particular, they were looking to classify microstructures as either dendritic or non dendritic. Chowdhury employed 8 different feature identification methods for a dataset of images. Classification was performed using an ensemble of ML techniques including support vector machines (SVM), Naïve Bayes, nearest neighbor, and a committee of the three previous classification methods [114]. Chowdhury's wide approach to image classification compared the predictive ability of all combinations of feature identification and classification methods, achieving classification accuracies above 90%.

It would be overly burdensome to lay out every *possible* application of computer vision in additive manufacturing. Efforts are underway across materials science

to implement computer vision for the automation of materials classification. Rather, the authors would like to refer the reader to reviews on the subject of computer vision for materials science, as well as open libraries. The hope is that readers will discover the many possible uses of computer vision and begin applying methods to their own AM problems.

## IV. LEARNING FROM THE PAST: MOVING TOWARDS DATABASE-DRIVEN DESIGN OF ADDITIVE

The genesis of this review article was motivated by successes of material scientists in applying data-driven methods over the past three decades. The scientific and engineering process of studying AM has developed similarly to other fields in materials science; particularly, materials design through thin film deposition and high throughput density functional theory have applied machine learning to find new materials with new properties. Materials science investigations in these regimes are focused around combinatorial screening of manufacturing or modeling inputs to search for optimized properties. As such, there is a good deal of information to be learned from these previous investigations which can be applied to ML of AM.

By the early 1990s, specifically engineered materials had become foundational in the modern technological world as more and more applications relied on advances in solid-state chemistry and physics. High-$T_C$ superconducting electrodes, high–energy density lithium-ion batteries, higher efficiency solid-state silicon transistors, and highly efficient photosensitive detectors and energy converters were all being sought. Improvements in the properties of commonly used materials would be advantageous for the scientific and engineering community.

At the same time, high-throughput material synthesis methods had also become a reality. Various forms of manufacturing and synthesis techniques allowed scientists and engineers to rapidly produce a span of materials at different stoichiometries with different precursors, different crystalline structures, and different properties. Chemical vapor deposition, metallorganic chemical vapor deposition, physical vapor deposition, and atomic layer deposition, among other techniques, became commonplace for the manufacturing of sensors, batteries, photovoltaics, electronics, and the like [31, 115–117]. Furthermore, electron microscopy and lab-scale X-ray diffraction techniques had matured to become routine measurements. Materials scientists now had the ability to quickly manufacture a range of different samples, then characterize their phases, structures, and properties faster than ever. Even further, computational materials science techniques matured during this time. Packages such as the Vienna Ab initio Simulation Package (VASP) were being used across the field for DFT simulations of material and chemical systems. As long as a material system was not too complicated, it could be accurately modeled in VASP. Materials scientists could manufacture and study materials systems at a wider and deeper scope than ever before. The problem, however, was *where* in the design space to manufacture.

Consider just the subset of all binary alloys: the number of different combinations of elements and stable structures has not yet been fully explored, even though scientists understand many of the underlying physical phenomena [118]. An AB metal crystal with a unit cell containing 10 atoms has $10^{11}$ different combinations [23]. It is possible to provide rigorous, fundamental quantum mechanical descriptions of atomic lattices for binary alloys, but it is not possible, in general, to design that rigorous description to have a specific set of desired properties. For example, scientists can readily predict the properties of a binary alloy and describe its formation; however, there are many possible binary alloys. It is often not a matter of a lack of understanding, but rather picking the right place to look.

In additive manufacturing, the same is true. High-fidelity computational models of AM exist and are widespread in literature [83]. The number of different manufacturing processes and alloys being developed is also widespread. If a scientist starts with an alloy, a deposition process, and manufacturing conditions, then it is reasonable to create an accurate computational model that will predict the final properties. The process of modeling and experimental verification can be time-intensive and cannot necessarily suggest *better* conditions under which to manufacture, however. Searching combinatorial spaces manually is currently the best manner to find local maxima of quality.

The ab initio process relies on having accessible methods of measuring a given material property. This typically means that the property can be quantified from a first-principles computation or an experimental measurement. The point is, the property itself needs to be stated in terms of measurable, mathematically expressible quantities. In the high-throughput ab initio materials science community, researchers often discuss the use of descriptors, which are chosen to screen materials for their properties. The very first ab initio investigations studied descriptors that were directly calculated from a first-principles approach; that is, they could be calculated from thermodynamic principles or by numerically solving the Schrödinger equation for a crystal.

For example, the superconducting electrode community has been searching for high-$T_C$ superconductors since the discovery of the phenomenon. In 1995, Xiang et al. at Lawrence Berkeley National Laboratory proposed the idea of combinatorial materials science for discovering such a superconductor [119]. Starting with seven binary compound precursors, they used an RF magnetron sputtering gun to synthesize 128 copper-oxide thin films that might exhibit superconducting properties. Xiang's group then characterized the superconducting properties at various composition points on the thin film substrates.

These measurements were related back to manufacturing conditions. Since the thin film had continuously varying composition based on manufacturing conditions, they were able to observe how the property varied over composition and structures. The results of their research were two new superconducting films, with $T_C$ of 80 and 90 K, respectively. This research was an early demonstration of the success of high-throughput methods.

Xiang et al. found that novel material systems can be screened from ab initio calculations of their properties with reasonable certainty. That is, their first-principles calculations matched well with the high-throughput characterization results. With this knowledge, scientists can design tests in a more informed manner without having to manufacture parts that definitely will not exhibit the properties of interest. The material systems in the study that seemed promising—meaning they exhibited desired properties, were stable in a solid state, or (ideally) both—were manufactured and characterized. High-throughput synthesis, computation, and characterization proved to be a feasible process by which to explore the structure-property space of superconductors. Following the study by Xiang et al., research teams across the globe began to investigate different niches of unexplored materials territory.

Initially, high-throughput ab initio experimentation like that of Xiang et al. was a common method of studying structure-property relationships. Investigations of these types are motivated by a logical thread of reasoning: if it is known that one stable crystal structure exhibits a desired property, then slightly altered crystal structures might exhibit the same—or better—properties. A scientist can manufacture a number of similar structures with slightly altered compositions, characterize their properties, then comment on trends in the composition-structure-property relationships. Fields whose materials are manufactured through chemical or physical deposition are especially suited for this type of research [32]. Common deposition techniques have enough degrees of freedom to allow for continuous composition variation within a single sample, which allows for continuous mapping of composition-structure-property relationships [120–122].

The same is true in additive manufacturing. It is expected that parts manufactured with identical geometries at identical conditions should demonstrate the same properties analogously to crystal structures and material properties. To reproduce mechanical properties like strength, ductility, hardness, etc. in AM, it must be determined *where* in the design space they occur. Following observation and characterization, the necessary design parameters can be adjusted to adopt those properties for other geometries, compositions, processes, or more. Operators can precisely control heat source parameters and energy density to produce a desired property. This is how the materials science community has conducted engineering research for a long time.

Computational high-throughput DFT studies also demonstrated success in the early 2000s. Computational ab initio has proven to accurately reproduce reality in many regimes when compared to experimental results [123]. While research groups may have access to rapid deposition techniques, it is expensive to manufacture swaths of composition-structure space. It also takes considerable time to manufacture and characterize samples. Considering the vastness of unexplored material compositions, structures, and properties, high-throughput experimental techniques can become infeasible.

By the late 1990s and early 2000s, computing power and accessibility skyrocketed while price dropped drastically, affording the materials science community the opportunity to combinatorially fill out the materials design space. The computational materials community began to *start* with the high-throughput method in the materials design process, as opposed to starting purely from first-principles or domain knowledge. The high-throughput ab initio method found many successes in materials science, ranging from the high-$T_C$ superconducting community [124] to lithium-ion battery development [27, 125–127], novel alloy discovery [42, 128], thermoelectric materials [129, 130], and novel electronic and piezoeletric materials [30, 131–133]. For a review of ab initio solid-state physics, chemistry, and materials research through the early 2000s, see Ref. [134].

As more and more materials scientists adopted the high-throughput method, huge repositories of new data were being created, often the results of DFT calculations. Databases of information were generated and disseminated to the scientific public [19, 20, 28, 131, 135, 136]. Large repositories of information can quickly become overwhelming and the amount of time and resources required to comb through the results is unreasonable. Even an expert in materials science cannot be expected to decipher physical trends from millions of data entries consisting of different descriptors all obtained by different means from different studies. The same will be true for AM and computational models of AM processes.

With all this computation being done, scientists could now compare across materials domains. So many research groups around the world were generating materials data that it became important to analyze trends across studies, materials, and disciplines. As a result, large materials informatics databases began appearing in academia. While databases have been crucial to materials science for some time—think the International Crystal Structure Database or Linus Pauling Files—these new databases were interactive, collaborative, and ever-growing. The more that research teams conducted computational studies, the more that materials informatics grew.

The generation of databases that are accessible to the scientific public is a primary step on the roadmap of the Materials Genome Initiative [137]. In an effort to reduce the design time on material systems, programs like the Materials Project incorporate data taken from a wide range of simulation methods into an open-source,

accessible database. The Materials Project also features electronic, structural, and thermodynamic calculations of different materials as well as an automated workflow for doing DFT computations of material systems [20]. Interactive computational databases provide not only a repository for materials scientists to store data but also a way for new research teams to adopt computational materials approaches.

These databases provide an additional benefit absent from high-throughput studies of individual material systems. Typically, a single ab initio investigation will focus on a specific material system, property, or composition range. By having large databases of information, calculations can be applied that find trends across these different material systems; thus, scientists can discover more fundamental phenomena that relate to the field of materials science as a whole.

In generating databases, scientists are attempting to achieve several different goals. A primary goal is open access to material information and a framework for material discovery. Having multiple teams across the world work on the same problem through a common interface allows for a deeper understanding of the field as more and more perspectives are generated. Similarly, having many different teams across the world work on different problems means that the space of possible material structures or properties is being fully explored. After all, it is very likely that improved or new materials have compositions or structures never before observed. The materials science community is trying to disseminate information quickly and efficiently to improve the search for new materials. In fact, some pipelines for high-throughput computation and analysis have included consideration of publication timelines in their processes [138]. The goal is to improve the field as quickly as possible while still generating consistent, accurate, and well-researched content. The same approach is aptly suited for characterizing AM systems and parts. Research and development in additive manufacturing has progressed to a level where large amounts of experimental and computational data are being generated, in the same manner that occurred for materials science in the early 2000s. Adopting machine learning algorithms to the simulation and experimental validation of printing processes will advance scientific understanding of the physics of AM while also expediting the design, development, and qualification of additively manufactured materials.

## V. CONCLUSIONS


[1] S. A. Khairallah, A. T. Anderson, A. Rubenchik, and W. E. King, Acta Materialia **108**, 36 (2016).
[2] W. Tan, N. S. Bailey, and Y. C. Shin, Computational Materials Science **50**, 2573 (2011).
[3] J. Kundin, L. Mushongera, and H. Emmerich, Acta Materialia **95**, 343 (2015).
[4] D. Pal, N. Patil, K. Zeng, and B. Stucker, Journal of Manufacturing Science and Engineering **136**, 1 (2014).
[5] J. Beuth and N. Klingbeil, Journal of Materials: Laser Processing (2001).
[6] J. Quinlan, Machine Learning **1**, 81 (1986).
[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, Science **290** (2000).
[8] S. T. Roweis and L. K. Saul, Science **290** (2000).
[9] R. Bro and A. K. Smilde, Analytical Methods Tutorial Review **6** (2014).
[10] J. J. Grefenstette, IEEE Transactions of Systems, Man, and Cybernetics **16** (1986).
[11] C. Boley, S. Mitchell, A. Rubenchik, and S. Qu, Applied Optics **55**, 6496 (2016).
[12] J. Trapp, A. M. Rubenchik, G. Guss, and M. J. Matthews, Applied Materials Today **9**, 341 (2017).
[13] G. Bi, C. N. Sun, and A. Gasser, Journal of Materials Processing Technology **213**, 463 (2013).
[14] J. H. Martin, B. D. Yahata, J. M. Hundley, J. A. Mayer, T. A. Schaedler, and T. M. Pollock, Nature Letters **549**, 365 (2017).
[15] P. Collins, D. Brice, P. Samimi, I. Ghamarian, and H. Fraser, Annual Review of Materials Research **46**, 63 (2016).
[16] C. A. Brice, W. A. Tayon, J. A. Newman, M. V. Kral, C. Bishop, and A. Sokolova, Materials Characterization (2018).
[17] A. Wang, S. Song, Q. Huang, and F. Tsung, IEEE Transactions on Automation Science and Engineering **14**, 968 (2017).
[18] P. Villars, N. Onodera, and S. Iwata, Journal of Alloys and Compounds **279**, 1 (1998).
[19] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, Computational Materials Science **58**, 227 (2012).
[20] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Materials **1** (2013), 10.1063/1.4812323.
[21] A. Fransceschetti and A. Zunger, Letters to Nature **402** (1999).
[22] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nature Materials **5**, 641 (2006).
[23] A. R. Oganov and C. W. Glass, Journal of Chemical Physics **124** (2006), 10.1063/1.2210932.
[24] Y. Ikeda, Materials Transactions **38**, 771 (1997).
[25] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, and T. Lookman, Nature Scientifi Reports **8** (2018).
[26] D. Wu, Y. Tian, L. Zhang, Z. Wang, J. Sheng, W. Wang, K. Zhou, and L. Liu, Materials **11** (2018).
[27] S. Kirklin, B. Meredig, and C. Wolverton, Advanced Energy Materials **3**, 252 (2013).
[28] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson, and S. Curtarolo, ACS Combinatorial Science ,

382 (2011).
[29] D. Wolf, O. Buyevskaya, and M. Baerns, Applied Catalysis A: General **200**, 63 (2000).
[30] A. Roy, J. W. Bennett, K. M. Rabe, and D. Vanderbilt, Physical Review Letters **109**, 1 (2012), arXiv:1107.5078.
[31] G. Gilmer, H. Huang, and C. Roland, Computational Materials Science **12**, 354 (1998).
[32] H. Koinuma and I. Takeuchi, Nature Materials **3**, 429 (2004).
[33] J. Morris, D. Deaven, and K. Ho, Physical Review B **53**, R1740 (1996).
[34] K.-M. Ho, A. A. Shvartsburg, B. Pan, Z.-Y. Lu, C.-Z. Wang, J. G. Wacker, J. L. Fye, and M. F. Jarrold, Nature (London) **392**, 582 (1998).
[35] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, and J. K. Norskov, Physical Review Letters **88**, 2555061 (2002).
[36] D. P. Stucke and V. H. Crespi, Nano Letters **3**, 1183 (2003).
[37] G. L. W. Hart, V. Blum, M. J. Walorski, and A. Zunger, Nature Materials **4**, 391 (2005).
[38] A. Kulkarni, K. Krishnamurthy, S. Deshmukh, and R. Mishra, Materials Science and Engineering A **372**, 213 (2004).
[39] W. Xu, P. R.-D. del Castillo, and S. van der Zwaag, Philosophical Magazine **88**, 1825 (2008).
[40] S. Dudiy and A. Zunger, Physical Review Letters **97** (2006).
[41] S. M. Anijdan, A. Bahrami, H. M. Hosseini, and A. Shafyei, Materials and Design **27**, 605 (2006).
[42] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chemistry of Materials **22**, 3762 (2010).
[43] N. Chakraborti, International Materials Reviews **49**, 246 (2004).
[44] G. A. Landrum and H. Genin, Journal of Solid State Chemistry **176**, 587 (2003).
[45] A. Oliynyk, E. Antono, T. Sparks, L. Ghadbeigi, M. Gaultois, B. Meredig, and A. Mar, Chemistry of Materials **28**, 7324 (2016).
[46] A. Antonysamy, J. Meyer, and P. Prangell, Materials Characterization **84**, 153 (2013).
[47] G. Strano, L. Hao, R. M. Everson, and K. E. Evans, Journal of Material Processing Technology **21**, 589 (2013).
[48] S. Bontha, N. W. Klingbeil, P. A. Kobryn, and H. L. Fraser, Materials Science and Engineering A **513-514**, 311 (2009).
[49] P. Nie, O. Ojo, and Z. Li, Acta Materialia **77**, 85 (2014).
[50] Y. Chen, F. Lu, K. Zhang, P. Nie, S. R. E. Hosseini, K. Feng, and Z. Li, Journal of Alloys and Compounds **670**, 312 (2016).
[51] B. Baufeld, E. Brandl, and O. Van Der Biest, Journal of Materials Processing Technology **211**, 1146 (2011).
[52] S. Bontha, N. W. Klingbeil, P. A. Kobryn, and H. L. Fraser, Journal of Materials Processing Technology **178**, 135 (2006).
[53] Y. Li and D. Gu, Materials and Design **63**, 856 (2014).
[54] J. Cherry, H. M. Davies, S. Mehmood, N. P. Lavery, S. G. R. Brown, and J. Sienz, International Journal of Advanced Manufacturing Technology **76**, 869 (2015).
[55] Q. Jia and D. Gu, Journal of Alloys and Compounds **585**, 713 (2014).
[56] J. Delgado, J. Ciurana, and C. A. Rodriguez, International Journal of Advanced Manufacturing Technology **60**, 601 (2012).
[57] A. M. Khorasani, I. Gibson, U. S. Awan, and A. Ghaderi, Additive Manufacturing (2018).
[58] A. S. Wu, D. W. Brown, M. Kumar, G. F. Gallegos, and W. E. King, Metallurgical and Materials Transactions A: Physical Metallurgy and Materials Science **45**, 6260 (2014).
[59] E. R. Denlinger, J. C. Heigel, P. Michaleris, and T. Palmer, Journal of Materials Processing Technology **215**, 123 (2015).
[60] C. E. Shannon, The Bell System Technical Journal **27**, 379 (1948), arXiv:9411012 [chao-dyn].
[61] J. Ling, M. Hutchinson, E. Antono, B. DeCost, E. A. Holm, and B. Meredig, Materials Discovery **10**, 19 (2017).
[62] B. Efron, Journal of the Royal Statistical Society Series B (Methodological) , 83 (1992).
[63] B. Efron, Journal of the American Statistical Association **109**, 991 (2014).
[64] S. Wager, T. Hastie, and B. Efron, Journal of Machine Learning Research **15**, 1625 (2014), arXiv:1311.4555v2.
[65] M. Gaultois, A. Oliynyk, A. Mar, T. Sparks, G. Mulholland, and B. Meredig, APL Materials **4**, 053213 (2016).
[66] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Phys. Rev. X **4**, 011019 (2014).
[67] C. Kim, G. Pilania, and R. Ramprasad, Chemistry of Materials **28**, 1304 (2016).
[68] S. Shan and G. G. Wang, Structural and Multidisciplinary Optimization **41**, 219 (2010).
[69] A. T. Gaynor and J. K. Guest, Structural and Multidisciplinary Optimization **54**, 1157 (2016).
[70] M. Langelaar, Additive Manufacturing **12**, 60 (2016).
[71] M. Langelaar, Structural and Multidisciplinary Optimization **55**, 871 (2017).
[72] T. Zegard and G. H. Paulino, Structural Multidisciplinary Optimization **53**, 175 (2016).
[73] E. Toyserkani, A. Khajepour, and S. Corbin, Optics and Lasers in Engineering **41**, 849 (2004).
[74] V. Manvatkar, A. De, and T. DebRoy, Journal of Applied Physics **116** (2014).
[75] D. Dai and D. Gu, Materials and Design **55**, 482 (2014).
[76] L.-Q. Chen, Annual Review of Materials Research **32**, 113 (2002).
[77] X. Gong and K. Chou, Journal of Materials **67**, 1176 (2015).
[78] S. Sahoo and K. Chou, Additive Manufacturing **9**, 14 (2016).
[79] P. Michaleris, Finite Elements in Analysis and Design **86**, 51 (2014).
[80] J. Ding, P. Colegrove, J. Mehnen, S. Ganguly, P. S. Almeida, F. Wang, and S. Williams, Computational Materials Science **50**, 3315 (2011).
[81] N. Raghavan, R. Dehoff, S. Pannala, S. Simunovic, M. Kirka, J. Turner, N. Carlson, and S. S. Babu, Acta Materialia **112**, 303 (2016).
[82] R. Martukanitz, P. Michaleris, T. Palmer, T. DebRoy, Z.-K. Liu, R. Otis, T. W. Heo, and L.-Q. Chen, Addtive Manufacturing **1**, 52 (2014).
[83] M. Francois, A. Sun, W. King, N. Henson, D. Tourret, C. Bronkhorst, N. Carlson, C. Newman, T. Haut, J. Bakosi, J. Gibbs, V. Livescu, S. Vander Wiel, A. Clarke, M. Schraad, T. Blacker, H. Lim, T. Rodgers, S. Owen, F. Abdeljawad, J. Madison, A. Anderson, J.-L. Fattebert, R. Ferencz, N. Hodge, S. Khairallah, and

O. Walton, Current Opinion in Solid State and Materials Science , 1 (2017).
[84] W. King, A. Anderson, R. Ferencz, N. Hodge, C. Kamath, and S. Khairallah, Materials Science and Technology **31** (2015).
[85] M. Gouge, P. Michaleris, E. Denlinger, and J. Irwin, *Thermo-Mechanical Modeling of Additive Manfuacturing, Chapter 2: The Finite Element Method for the Thermo-Mechanical Modeling of Additive Manufacturing Processes* (Elsevier Inc, 2018).
[86] J. A. Flores-Livas, A. Sanna, and S. Goedecker, Novel Superconducting Materials **3**, 6 (2017).
[87] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical Review Letters **208**, 1 (2011), arXiv:1109.2618.
[88] J. C. Snyder, M. Rupp, K. Hansen, K. R. Müller, and K. Burke, Physical Review Letters **108**, 1 (2012), arXiv:1112.5441.
[89] C. Kamath, International Journal of Advanced Manufacturing Technology **10** (2016).
[90] R. R. Dehoff, M. M. Kirka, W. J. Sames, H. Bilheux, A. S. Tremsin, L. E. Lowe, and S. S. Babu, Materials Science and Technology **31** (2015).
[91] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Physical Review Letters **91**, 135503 (2003), arXiv:0307262 [cond-mat].
[92] K. Rajan, C. Suh, and P. F. Mendez, Statistical Analysis and Data Mining **1**, 361 (2009).
[93] G. Tapia, S. A. Khairallah, M. Matthews, and W. E. King, International Journal of Advanced Manufacturing Technology **10** (2017).
[94] W. E. King, H. D. Barth, V. M. Castillo, G. F. Gallegos, J. W. Gibbs, D. E. Hahn, C. Kamath, and A. M. Rubenchik, Journals of Materials Processing Technology **214**, 2915 (2014).
[95] M. A. Bessa, R. Bostanabad, Z. Liu, A. Hu, D. W. Apley, C. Brinson, W. Chen, and W. K. Liu, Computer Methods in Applied Mechanics and Engineering **320**, 633 (2017).
[96] Y. Jin and B. Sendhoff, IEEE Transactions of Systems, Man, and Cybernetics - Parc C: Applications and Reviews **38** (2008).
[97] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Physical Review B - Condensed Matter and Materials Physics **89**, 1 (2014).
[98] S. Berumen, F. Bechmann, S. Lindner, J.-P. Kruth, and T. Craeghs, Physics Procedia **5**, 617 (2010).
[99] G. Tapia and A. Elwany, Journal of Manufacturing Science and Engineering **136**, 060801 (2014).
[100] M. Mani, B. M. Lane, M. A. Donmez, S. C. Feng, and S. P. Moylan, International Journal of Production Research **55** (2017).
[101] J. T. McKeown, K. Zweiacker, C. Liu, D. R. Coughlin, A. J. Clarke, J. K. Baldwin, J. W. Gibbs, J. D. Roehling, S. D. Imhoff, P. J. Gibbs, D. Tourret, J. M. Wiezorek, and G. H. Campbell, JOM **68** (2016).
[102] U. S. Bertoli, G. Guss, S. Wu, M. J. Matthews, and J. M. Schoenung, Materials and Design **135**, 385 (2017).
[103] J. Raplee, A. Plotkowski, M. M. Kirka, R. Dinwiddie, A. Okello, R. R. Dehoff, and S. S. Babu, Scientific Reports **7**, 1 (2017).
[104] D. G. Lowe, International Journal of Computer Vision **60**, 91 (2004).
[105] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, Computer Vision and Image Understanding **110**, 346 (2008).
[106] B. Yuan, G. M. Guss, A. C. Wilson, S. P. Hau-Riege, P. J. DePond, S. McMains, M. J. Matthews, and B. Giera, Adv. Mater. Technol. , 1 (2018).
[107] L. Scime and J. Beuth, Additive Manufacturing **24**, 273 (2018).
[108] B. L. DeCost and E. A. Holm, Computational Materials Science **126**, 438 (2017).
[109] J. Zhou, Y. Zhang, and J. Chen, Journal of Manufacturing Science and Engineering **131**, 1 (2009).
[110] C. Boley, S. Khairallah, and A. Rubenchik, Applied Optics **54**, 2477 (2015).
[111] B. L. DeCost, H. Jain, A. D. Rollett, and E. A. Holm, Jom **69**, 456 (2017).
[112] B. L. Decost and E. A. Holm, Computational Materials Science **110**, 126 (2015).
[113] B. L. DeCost, T. Francis, and E. A. Holm, Acta Materialia **133**, 30 (2017), arXiv:1702.01117.
[114] A. Chowdhury, E. Kautz, B. Yener, and D. Lewis, Computational Materials Science **123**, 176 (2016).
[115] M. J. Hampden-Smith and T. T. Kodas, Chemical Vapor Deposition **1** (1995).
[116] B. Mercey, P. A. Salvador, W. Prellier, T.-D. Doan, J. Wolfman, J.-F. Hamet, M. Hervieu, and B. Raveau, Journal of Materials Chemistry **9**, 233 (1999).
[117] D. B. Mitzi, Chemical Materials **13**, 3283 (2001).
[118] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Scientific Reports **3**, 2810 (2013).
[119] X. D. Xiang, X. Sun, G. Briceno, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, and P. G. Schultz, Science **268**, 1738 (1995).
[120] C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, and X. Li, Review of Scientific Instruments **78** (2007), 10.1063/1.2755487.
[121] C. J. Long, D. Bunker, X. Li, V. L. Karen, and I. Takeuchi, Review of Scientific Instruments **80** (2009), 10.1063/1.3216809.
[122] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long, and I. Takeuchi, Scientific Reports **4**, 6367 (2015).
[123] S. Curtarolo, D. Morgan, and G. Ceder, Calphad: Computer Coupling of Phase Diagrams and Thermochemistry **29**, 163 (2005), arXiv:0502465 [cond-mat].
[124] A. N. Kolmogorov and S. Curtarolo, Physical Review B - Condensed Matter and Materials Physics **73**, 1 (2006), arXiv:0603304 [cond-mat].
[125] H. Chen, G. Hautier, A. Jain, C. Moore, B. Kang, R. Doe, L. Wu, Y. Zhu, Y. Tang, and G. Ceder, Chemistry of Materials **24**, 2009 (2012).
[126] G. Hautier, A. Jain, T. Mueller, C. Moore, S. P. Ong, and G. Ceder, Chemistry of Materials **25**, 2064 (2013).
[127] K. Kang, Y. S. Meng, J. Breger, C. P. Grey, and G. Ceder, Science **311**, 977 (2006).
[128] C. V. Ciobanu, D. T. Tambe, and V. B. Shenoy, Surface Science **582**, 145 (2005).
[129] S. Wang, Z. Wang, W. Setyawan, N. Mingo, and S. Curtarolo, Physical Review X **1**, 1 (2011).
[130] J. Yan, P. Gorai, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, V. Stevanovic, and E. S. Toberer, Energy & Environmental Science **8**, 983 (2015).
[131] W. Setyawan and S. Curtarolo, Computational Materials Science **49**, 299 (2010), arXiv:1004.2974.

[132] J. W. Bennett, K. F. Garrity, K. M. Rabe, and D. Vanderbilt, Physical Review Letters **109**, 1 (2012), arXiv:1206.4732v1.

[133] M. de Jong, W. Chen, H. Geerlings, M. Asta, and K. A. Persson, Scientific Data **2**, 150053 (2015).

[134] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature Materials **12**, 191 (2013).

[135] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan, Computational Materials Science **58**, 218 (2012), arXiv:1308.5715.

[136] A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, Computational Materials Science **50**, 2295 (2011).

[137] J. J. De Pablo, B. Jones, C. L. Kovacs, V. Ozolins, and A. P. Ramirez, Current Opinion in Solid State and Materials Science **18**, 99 (2014), arXiv:arXiv:1011.1669v3.

[138] I. Foster, R. Ananthakrisnan, B. Blaiszik, K. Chard, R. Osborn, S. Tuecke, M. Wilde, and J. Wozniak, Big Data and HPC (2015).