INSTITUTE FOR DEFENSE ANALYSES

Data Science for Materials in Extreme Environments

Jessica G. Swallow

September 2018

Approved for public release;
distribution is unlimited.

IDA Document D-9301

Log: H 18-000430

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, IDA Central Research Project C2229. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information:

Jessica G. Swallow, Project Leader
jswallow@ida.org, 703-578-2892

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, 703-578-2800

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-9301

Data Science for Materials in Extreme Environments

Jessica G. Swallow

Executive Summary

Many Department of Defense (DoD) systems experience extreme environments during regular service life, including very high or low temperatures, corrosive environments, radiation, and high strain rates. Extreme environments can have adverse effects on material properties, such as changes in strength or toughness, by affecting composition and microstructure. Designing, modeling, and managing materials in extreme environments is an important mission for the DoD because it affects platform lifetime, reliability, maintenance, and performance specifications.

A challenge of modeling and designing materials for extreme environments is the complexity of processes that lead to failure. Both composition and microstructure may vary with time upon exposure to extreme environmental conditions, and the relationships between these variables and materials properties that dictate performance are not always well understood or easy to elucidate. It is difficult to independently control various aspects of microstructure, and the mechanisms that determine microstructural evolution are also often poorly understood or difficult to model. Data on materials in extreme environments can be painstaking to collect, often requiring specialized apparatus and long time periods. This combination of factors makes data science methods that employ soft-modeling or statistical learning approaches particularly attractive because these can identify hidden trends in data for further investigation, extract key variables, and assist with visualizing complex relationships that may be difficult to establish deductively. Such methods shine when empirical models are unavailable or of limited applicability.

Specifically, materials scientists seek to identify processing-structure-property (PSP) linkages that can explain how to engineer materials to achieve particular desired properties. In recent years, efforts applying data science approaches to this goal have met with some success in crystal structure identification, property prediction, and microstructure description. In particular, databases of computed properties based on density functional theory and thermodynamic calculations have led to a number of discoveries of new materials for applications ranging from transparent conducting oxides to magnets and Li-ion battery materials. In few cases have these methods, or any data science methods, been applied to problems specific to materials in extreme environments. One important reason for this is that material failures due to extreme environments are often mechanical and therefore closely connected to microstructural variables, which in turn depend on processing conditions and are thus much harder to predict from first principles. Other barriers include a lack of available data in consistent formats and limited awareness of or access to statistical learning methods and tools.

The lack of data can be attributed to several important challenges for building databases for materials in extreme environments. Firstly, materials are generally characterized on many length

scales, and microstructures may include many different types of features or defects, meaning that a comprehensive description of material microstructure and composition needs to be flexible enough to incorporate an enormous variety of data even for a single material family. Secondly, significant variety exists in the characterization methods chosen by researchers, such that data may come with substantial uncertainty. The culture of reporting data via journal article that is prevalent for materials researchers is also not conducive to straightforward database building. Partly, this is because it is essential to capture metadata such as experimental and processing conditions in addition to microstructural information to build databases that are useful for establishing PSP linkages. Few existing databases meet all of these parameters.

Many tools are now available to help build an infrastructure of materials data science and materials informatics. Several materials databases do exist (though rarely specific to materials in extreme environments), which may contain computational or experimental data, or both. There are also a number of materials resource and data repositories available, as well as several efforts to build open-source analysis, visualization, and e-collaboration tools for materials researchers. Few of these tools are specific to materials in extreme environments, and many are not well-known to the DoD materials testing and research community. Some statistical learning tools of particular interest for materials scientists are those that permit multi-label classification, fuzzy logic, or automated, computational image processing and analysis.

Interviews with researchers and materials testing experts in government labs focused on materials in extreme environments revealed that these groups have interest in using such tools and building databases to preserve important results. Despite this interest, such actions have not been adopted because of limited access to up-to-date software and lack of institutional support for extensive data entry programs. To prevent the loss of useful, expensive data through personnel turnover and failure to document or store, we recommend that organizations collecting such data implement policy changes that will encourage researchers to store their test data in accessible, curated repositories. We also recommend a full review of data management protocols in DoD materials testing and characterization laboratories to identify gaps and best practices.

There are a number of DoD-relevant material classes in which a curated and accessible database would be useful, including high-entropy materials, corrosion-resistant materials, radiation-resistant materials, high-temperature structural materials, and thermal barrier coatings. In particular, due to the relatively young state of the field and the small size of the existing community, high entropy materials may represent a flagship area for establishing standardized data reporting and curation habits. Corrosion and radiation-resistant materials researchers may benefit from computational image analysis methods attached to existing databases of material surveillance data. High temperature materials and thermal barrier coatings of interest for hypersonic applications may benefit from automated image analysis and database building for process control and quality assurance. Databases of computationally-generated microstructures could also be of interest for building models that explain the relationships between complex microstructural variables and material properties.

Databases should be structured to assist materials data scientists in designing new analyses. Potential users need to understand the nature of data contained in a database, including its quantity (how many data points?) and quality (how reliable are the data?), the types of materials covered (compositions, processing methods, etc.), and the types of data contained (e.g. images, spectra, measured properties, compositions). Failure to adequately describe database contents in a public and accessible way is a significant barrier to researchers who might use such data. If the goal is to make a useful resource, it is essential to provide detailed and well-advertised information about database contents.

Case studies on limited sets of experimental data identified additional challenges and details of value for building resources for data science for materials in extreme environments. Analysis of composition and crystal structure information for a table of 110 high entropy alloys demonstrates that support vector machines, decision trees, and random forests are all adequate methods for conducting multi-label crystal structure classification in this class of materials. However, more data are needed to build robust classifiers that can accommodate a variety of intermetallic phases. This is an example of the type of classification that could be conducted starting from computationally-generated data, but must be validated with experimental data and an awareness of processing methods. Regression on data of composition and exposure conditions of irradiated FeCrAl alloys was able to predict yield strength with ~15% error, but results were difficult to interpret due to lack of microstructural information and variable correlation. Incorporating more physical intuition and microstructural variables into the model-building process would likely yield improved results. Finally, computational image analysis was applied to a number of optical micrographs of grey cast iron. This analysis led to the essential conclusion that automated imaging pipelines must be well-documented and validated in order to quantify uncertainty and delineate limits of applicability. With adequate documentation and validation, automated image analysis can significantly increase the amount of data that a single researcher can process at a time. However, to extend the reach of such tools, they need to be incorporated into the materials training pipeline. Universities should cover both quantitative image analysis and general statistical learning methods in materials science curricula for both undergraduate and graduate students. Furthermore, university courses and libraries can help publicize the existence of various materials data science tools, including databases, repositories, and software, to improve the awareness and technical skills of the materials science workforce.

Contents

1. Introduction: Data Science and Materials in Extreme Environments

Extreme environments are ubiquitous in military systems, which during service may experience chemical attack, mechanical stress, irradiation, extremes in temperature and pressure, or combinations of these conditions. Materials determine the performance of platforms like hypersonic aircraft, naval vessels, and nuclear reactors, defining maximum and minimum allowable usage temperatures, limiting service lifetime, and dictating maintenance requirements. For example, the permissible speed of a hypersonic aircraft is directly limited by the thermal properties of the leading edge and thermal protection system in atmosphere. (Sziroczak and Smith 2016) Meanwhile, many platforms will experience cracks and mechanical failures after repeated or long-term exposure to corrosive environments or radiation. Military platforms may also be subject to extreme stimuli including high strain rates, electric fields, or chemical hazards. Development, design, and modeling of materials for extreme environments is therefore of significant value to the Department of Defense (DoD).

Extreme environments can be damaging to material properties, especially during long-term or cyclic exposure. For example, high temperatures promote surface reactions such as oxidation, enable microstructure changes such as grain growth or *coarsening*, and facilitate phase separation through interatomic diffusion. Similarly, corrosion causes gradual changes in material microstructure, affecting both composition and mechanical properties of materials, and may be highly dependent on specific environment descriptors. Radiation causes atomic displacements in materials that can lead to void formation, phase separation, and embrittlement. Cold temperatures can likewise cause embrittlement, while at the same time arresting transport properties such as diffusivity that may be important to material function (e.g., for electrochemical devices such as batteries).

It is difficult to characterize many of these material degradation mechanisms. Not only are the physical processes involved often slow (a typical corrosion rate is measured in mm/year), but they may also be poorly understood. In addition, there are often numerous variables that may contribute to a particular property of interest, including test conditions, initial composition, and material microstructure and/or processing. Many of these variables are difficult to control independently, and characterization and quantification may be painstaking as well, especially for microstructural variables. However, designing materials to withstand and function properly in extreme environments is required to expand the capabilities of technologies ranging from nuclear reactors to hypersonic aircraft. Data collected in the area of materials for extreme environments is difficult to come by and therefore extremely valuable.

Recently, the concept of data science has expanded as a technical discipline that touches nearly every other type of science, ranging from market analysis, to bioinformatics, to artificial intelligence. Data science encompasses data management (e.g., data storage, collection, tracking, and sharing), data analytics (e.g., statistical inference, computational algorithms), and e-collaboration (Kalidindi and Graef 2015). Data science provides tools to address specific technical challenges, such as cluster analysis, classification, or prediction. The methods used may be supervised or unsupervised, and will generally be more or less successful depending on the character of the data being analyzed. The data in question can also take many forms—from images, to purchase histories, to experimental measurements and conditions. In the materials science community, images, experimentally measured properties (including composition), measurement conditions, and processing history are some of the most important types of data.

Data science methods are of interest to materials scientists primarily for identifying processing-structure-property relationships (Kalidindi and Graef 2015; Rajan 2015; Agrawal and Choudhary 2016). The ultimate goal of such analysis is to solve inverse design problems and to increase the amount of data-driven decision making in the materials design process. There are many challenges in place for realizing this vision, most notably the variety and complexity of materials data. Material characterization takes place over many length scales and can be conducted with a range of techniques that describe composition, microstructure, and materials properties. For example, a single metallic alloy may contain features including point defects, grains, dislocations, precipitates, twin boundaries, and pores, and each of these features will have a particular orientation, distribution, and associated strain or composition fields. These features are not static over time, especially in extreme environments that often promote morphology or composition change. However, they play an essential role determining the performance characteristics of materials, including mechanical and transport properties. For the materials community, describing such varied data in consistent, useful, and standard ways while making it accessible to those who might conduct useful analysis is a challenge. Because of such challenges, statistical learning approaches remain beyond the scope of much materials science research.

This report assesses the challenges and opportunities for applying statistical learning approaches to understanding materials in extreme environments, with particular focus on challenges of interest to the DoD. We begin in Chapter 2 by highlighting a few examples where data science has been successfully applied to materials problems, focusing on the methods used and why they are of particular value to the materials community.

Chapter 3 then reviews the status of available databases, data repositories, and open-access data science tools for materials science. Such resources are growing in number and availability through projects such as the Materials Genome Initiative (Jain et al. 2013), NIST's Materials Resource Registry (NIST n.d.), and Georgia Tech's MATIN (Georgia Tech n.d.). A list of data science resources and associated URLs can be found in Appendix B. Chapter 3 also covers the challenges for adapting or designing such tools for the special case of materials in extreme environments, including perspectives from a few experts in the field of extreme-environment

material characterization. The communities for extreme-environment materials research are often quite small, with highly specialized testing methods and data that preclude standardized, widely accessible forms. However, there are a few cases where subject-specific data have been compiled—these are highlighted with specific comments as to the utility of such databases, potential for expansion, and particular value to the DoD.

Chapter 4 identifies several target areas of value to the DoD that appear well-suited for a materials informatics or statistical learning approach. For each area, potential descriptor variables, target variables for optimization, analysis and data production methods, and obstacles to implementation are described.

Chapter 5 presents three case studies where statistical learning approaches were applied to small available materials data sets. These case studies include a classification problem involving high entropy alloys, a regression problem involving radiation resistance in FeCrAl alloys, and an image analysis problem with data pulled from the ASM Microstructure Database. This chapter is primarily exploratory in nature, simply testing out a few methods on real data and describing the advantages and pitfalls of these approaches, including the role of missing data.

Finally, we conclude with a perspective on the path forward for the DoD to maximize the value of its hard-won data in understanding, modeling, and designing materials for extreme environments.

2. Data Science Methods in Materials Science: Representative Examples

In this section, we focus on only a few examples of data science or statistical learning methods that have been applied in materials research and appear promising for future work on materials in extreme environments. The purpose of this section is to illustrate ways in which these methods have been successful, rather than to fully review the topic. For more complete coverage of previous materials data science work, we refer the reader to a group of recently published comprehensive reviews (Kalidindi and Graef 2015; Rajan 2015; Agrawal and Choudhary 2016; Ramprasad et al. 2017).

A. Classification Methods

Classification is a common goal of statistical learning that may be applied to both ordered problems (e.g., ranking of susceptibility to a particular failure mechanism) and non-ordered problems (e.g., predicting crystal structure). In materials science, classification might also involve comparing an image to a known microstructure standard to determine class membership. This can be used as a kind of quality assurance in manufacturing or monitoring processes. Finally, we might use classification to determine whether an image contains certain features (e.g., dislocation loop, grain boundary, precipitate, pore), or to find certain features within an image. Such automated feature detection may also be extended to data collected by spectroscopic methods or other means. In fact, such feature detection algorithms, especially as applied to images, are general to many data-science focus areas, such as facial recognition or medical image analysis (Russ 2011).

A study by Kong et al. provides a model of data-driven classification as applied to materials science. In this study, crystal structure data for 840 intermetallic compounds were gathered and used to identify design rules based on classification trees (Kong et al. 2012). The authors chose descriptor variables that could be determined solely based on material composition (such as average number of valence electrons per atom, or average difference in atomic radii). They then constructed classification trees by using information entropy (a mathematical descriptor of uncertainty in data) as a partitioning metric, and used cross-validation to determine the optimal tree depth. The final tree could determine a subset of likely crystal structures (from a possible 34) based solely on input composition. The output of this tree allowed the authors to reduce computational effort for full-scale density functional theory simulations of unknown structures by ~85%. Moreover, because the classification-tree method maintains explanatory power, it provided physical intuition that would have been absent from something like a neural network.

It is important to note, however, that *crystal structure prediction* is a much more tractable challenge than *microstructure prediction*. In the absence of specific efforts to quench a material into a thermodynamically metastable state, crystal structure is heavily determined by thermodynamics, while microstructure is heavily dependent on processing conditions. Two samples with identical crystal structure can easily have significantly different material properties; this is due to variation in microstructure. Unless a material is intended to be used as a single crystal, designing materials for particular applications requires attention to microstructure and processing. In the context of extreme environments, processing can also include usage profiles, such as thermal history or radiation exposure. Regardless, thermodynamic predictions can still be quite valuable for materials in extreme environments; for example to predict which oxide scales are most likely to form as products of high temperature oxidation.[1]

Often, a single sample contains a variety of microstructural features or phases. Therefore, classification methods that allow for multiple labels are of particular interest to materials scientists, especially in cases where class membership may exhibit conditional relationships. For example, small adjustments in composition can shift thermodynamic equilibria in alloys between solid solution (a single-phase state) and phase-separated states. Rather than considering a phase-separated state as a separate class, one can consider it as a state in which two classes coexist. This can be useful from a mathematical perspective, because it prevents us from either ignoring multi-label data or creating unique classes for multi-label data that may be relatively sparse. Classification trees and support vector machines (SVMs) are statistical learning methods that are well-adapted to handling multi-label classification problems.

SVMs seek to maximize the margin between multiple classes based on so-called *supports*, which are the nearest points on either side of the boundary (Hastie, Tibshirani, and Friedman 2009). One can think of an SVM as a type of regularization that determines a boundary by penalizing points for being too close to the boundary, or for being on the wrong side of the boundary. SVMs can describe non-linear boundaries, and can be tuned for the softness of a boundary by adjusting the regularization parameter *C*. Furthermore, SVMs are easily adapted to multi-label problems by training classifiers using a one-vs-all or *cross-training* approach (Boutell et al. 2004). Multi-label data points are often on the edges of class regions, and therefore SVMs are well-suited to taking advantage of such data points. SVMs can also be used for prediction, for example to relate methane storage capacity to structural descriptors of metal-organic frameworks (Fernandez et al. 2013). Chapter 5.A describes a case study on high entropy alloys inspired by the work of Kong et al. that will explore the use of SVMs and decision trees for multi-label classification.

Although we did not cover it in the case studies, another useful concept that is important for classification problems in materials science is *fuzzy logic*. Algorithms that employ fuzzy logic allow for a certain amount of uncertainty in class membership, typically predicting *probability of*

[1] Based on conversation with Dr. James Smialek, NASA

class membership rather than binary class membership (Rajan 2015). They account for *indiscernability* in data—i.e., the situation where two points with identical descriptor variables can fall into separate, mutually exclusive classes (Jackson, Pawlak, and LeClair 1998; Boutell et al. 2004). While mathematically distinct from multi-label classification, fuzzy classification can use similar learning strategies (including SVMs and decision trees) (Boutell et al. 2004). Boundary cases in fuzzy logic classification tend to be related to the uncertainty surrounding decision boundaries, rather than coexistence of multiple classes. This is an appealing method for materials data analysis because predictor variable sets are often incomplete, and data may have significant uncertainty, especially when translated between different labs or experimental techniques. Indeed, the sharpness of boundaries between classes can itself be a useful output from data analysis.

B. Property Prediction

The second broad statistical learning goal in materials science is prediction. Although it is one of the most widely used prediction methods, regression, even linear regression, can be subtle to implement, especially when applied to high-dimensional data. Accounting for covariance in variables, selecting predictor variables, and checking that assumptions are fulfilled are all important steps in any regression analysis. When applied well, regression can provide both explanatory power and property prediction. When applied blindly, regression can be misleading or fail to capture useful information contained in existing data.

A study of high-temperature oxidation resistance conducted by NASA is an illustrative example of regression applied well in the area of materials for extreme environments (Smialek and Bonacuse 2016). High temperature oxidation resistance is of great concern for turbines, both for jet engines (Air Force) and for spacecraft (NASA). Over several decades, NASA collected data on weight loss due to oxidation and spalling of alloys and superalloys subjected to cyclic high-temperature oxidation conditions. (Smialek and Barrett 1997) The data from these experiments were stored in a database that researchers at NASA were able to analyze in detail in a number of studies using statistical learning and regression approaches (Smialek and Bonacuse 2016; Barrett 1997, 1992; Smialek et al. 2000; Barrett 2003).

In a 1992 study, the weight loss data were used to build a metric called an “attack parameter” K_a for Ni- and Co-base superalloys that accounted for both scale growth rate and spalling rate (Barrett 1992). Using multiple linear regression, Barrett constructed a model of high temperature-oxidation susceptibility as a function of composition that accounted for 85% of observed variance with 14 terms, and used that model to predict optimal compositions for this type of alloy. Barrett selected first order terms by stepwise selection at a specified significance level. For second order terms, Barrett used a partial subset analysis procedure. Later, in 2016, a new analysis of Ni-base superalloys was conducted using the same data. (Smialek and Bonacuse 2016) In that analysis, a logarithmic weight-gain transform was fit to compositional variables. Terms were selected by considering the Akaike Information Criterion (AIC) and design matrix condition numbers of models including those terms. The AIC and design matrix condition number, respectively, penalize

models with additional terms that add minimal improved *goodness of fit* and quantify the degree of covariance of model terms. The more recent model, which focused only on Ni-base superalloys, used 10 terms and accounted for 84% of data variance.

These studies illustrate several useful points for the use of regression in understanding complicated extreme-environment data. First, despite a space of over 100 potential variables to include in the models, the models were constructed with a small subset of these, and selections were made based on objective tools for variable importance quantification. Beyond simply conducting regression fits, the authors of both studies also did extensive analysis of the complex, often interdependent contributions of composition to the overall oxidation resistance of the alloys. Finally, although the data used for this study were collected more than 15 years ago, their inclusion in an accessible database within NASA enabled useful, new analysis as enhanced computational power became available.

The two analyses also frame the data differently; one focused on attack parameter, the other on weight change. In fact, the attack parameter could have been analyzed completely differently. This type of parameter has been used to rank alloys into classes of oxidation susceptibility—such ranking schemes are perfect for ranked classification problems such as ordered logistic regression. (Barrett 1997) Where data are not of sufficient quality for quantitative property prediction, such classification methods can be a powerful alternative analysis option.

We also note that the database used for the above study did not include microstructural information. In high temperature oxidation processes, the type and proportion of scale growth depends heavily on thermodynamic variables. Therefore, composition alone can be a useful predictor of oxidation resistance, because composition influences the shift of equilibrium toward protective scales, like alumina, and away from detrimental scales, such as those containing Ti and Nb. Knowledge of which elements were helpful or harmful to oxidation resistance was directly derived from detailed analysis of the contributions of various regression terms (Smialek and Bonacuse 2016). That said, some variance in experimental results could not be explained solely through compositional analysis, and the protective value of certain elements (e.g., Ta) could not be adequately explained through the formation of a known protective scale. It is here that microstructural information (such as initial grain sizes, phase fractions, or phase morphologies) could be useful, because it can provide the missing links that explain why certain elements are important, or that variance is due more to kinetic rather than thermodynamic effects. We shall see some consequences of such missing information in chapter 5.B, where we apply linear regression with regularized variable selection to data on radiation resistance in FeCrAl alloys.

C. Microstructure Description

A key challenge relevant to both prediction and classification in materials data science is quantifying microstructure. As described in the Introduction, microstructure may include many variables and length scales, and may be characterized in many ways, including myriad varieties of microscopy, acoustic methods, and tomographic (3-D) methods. Microstructure description must

be *flexible* and *extensible*, and ideally it must also be quantitative. While we cannot expect any particular set of microstructural variables to be sufficient for all possible materials of interest, we might seek methods of quantifying microstructure that can be adapted to any particular problem.

The work of Kalidindi et al. gives us a convenient strategy to address this challenge using n-point statistics (Kalidindi, Niezgoda, and Salem 2011). N-point statistics, roughly speaking, describe the probability that two local states will be found at the end points of some n-dimensional volume. A single image may contain only two local states of interest, such as two different phases, or it may contain many, such as a range of grain orientations. The n-point statistic provides a mechanism for quantifying the arrangement of these states that is flexible and extensible enough to cover a range of length scales, materials, and local states of interest. Two- and three- point statistics are particularly useful due to the ease of estimation from images using Fast Fourier Transforms (FFT) and their ability in some cases to actually regenerate a microstructure from compact storage forms (Kalidindi, Niezgoda, and Salem 2011; Niezgoda, Kanjarla, and Kalidindi 2013). One can think of this as a generalized way of quantifying feature distribution or orientation functions.

When combined with the dimensionality-reduction method known as principal component analysis (PCA), Kalidindi et al. demonstrated this approach for unsupervised and supervised clustering and classification for both experimental Ti alloy micrographs and computationally generated microstructure datasets (Kalidindi, Niezgoda, and Salem 2011; Niezgoda, Kanjarla, and Kalidindi 2013). The addition of known material properties (e.g., stiffness) or processing information (e.g., annealing time and temperature) to instances of quantified microstructure then facilitated generation of processing-microstructure-property linkages. Specifically, these authors were able to link variation in microstructure to variation in a material property of interest (stiffness) (Niezgoda, Kanjarla, and Kalidindi 2013). While PCA and n-point statistics do abstract the representation of material microstructure, they also provide a convenient means of comparing and quantifying otherwise highly complicated data that may include feature shape, size, orientation, or distribution.

The open source PyMKS software package provides tools for applying n-point statistics to microstructural images and studying PSP linkages (Brough, Wheeler, and Kalidindi 2017). Such tools are readily applied to computationally-generated microstructures (which can also be built using PyMKS), as the individual states of interest are typically labeled in the generation process. However, we will see in chapter 5.C that when used to quantify experimental images, the results of this method combined with PCA are heavily dependent on the choices made during the pre-processing steps of image filtering and segmentation. Therefore, for collaboration across multiple labs and data sets, this technique requires advanced consideration of which features are important, which data to store or share, and methods for standardizing image collection, filtering, and segmentation or quantifying the error of these processes.

3. Status of Available Materials Databases and Data Science Tools

Materials data science tools fall into several broad classes: databases, data repositories, and analysis tools. This chapter will briefly review available resources in each area, citing strengths, weaknesses, and application as appropriate. The challenges of hosting, maintaining, and curating such databases are well-described in multiple reviews (Tenopir et al. 2011; Kalidindi and Graef 2015; DMMI 2014). The resources mentioned here are listed in Appendix A, but the highlighted resource registries should provide more comprehensive lists.

A. Databases

Databases for material science may contain computational data, experimental data, or both, and may be commercial, open access, or limited access. This discussion will not include databases of published journal articles, because these do not allow for straightforward data extraction for large-scale analysis of PSP linkages. The largest open-source databases tend to be those containing computationally-generated predictions of crystal structure, electronic structure, thermodynamic properties, and phase diagrams. Such databases include the Materials Project, Open Quantum, and AFLOW, and each includes tools for running workflows to add new data to the database (Jain et al. 2013; Saal et al. 2013; Curtarolo et al. 2012). These databases support a materials design approach whereby many potential compositions and/or crystal structures are screened for computable properties of interest (e.g., band gap, effective mass of charge carriers, Curie temperature), and the most promising candidates are then synthesized in the laboratory for full testing. Such an approach has been applied to screening for transparent conducting oxides (Hautier et al. 2013), new magnetic materials (Sanvito et al. 2017), and Li-ion battery materials (Saal et al. 2013; Qu et al. 2015).

The beauty of such an approach is that it allows investigators to generate the data they need without any laboratory equipment. The drawback is that it provides limited mechanisms for considering the effects of microstructure and processing. It turns out that the greatest challenges of materials in extreme environments are those related to changes in material properties over time. Material property drift results from both compositional and microstructural changes that are often stimulated by exposure to factors such as elevated temperatures, radiation, or chemical attack. In such a context, thermodynamic predictions are of limited value, because they do not account for processing history that may create kinetic barriers to achieving thermodynamic equilibrium. Instead, modeling that can capture the processes that cause microstructure or composition change

in response to stimuli like radiation and heat are needed, especially if these models can also relate such changes in composition and microstructure to properties like strength or toughness.

The second common type of materials database compiles experimentally measured materials properties. These may be established standards, such as those reported by NIST (Johnson 2010), data compiled from literature sources, as reported in the NIMS MatNavi Database (Xu, Yamazaki, and Villars 2011), or carefully curated collections of specific data, as in the International Crystal Structure Database, ICSD (Allmann and Hinek 2007). Such databases often contain composition, crystal structure, and property information, but there are few databases that also include microstructural details or detailed processing histories, and even fewer that include actual micrographs. The subscription-only ASM Micrograph database is an exception in that it contains thousands of micrographs for many different materials, including limited processing and compositional information (ASM International n.d.). However, even this example rarely provides micrographs covering multiple length scales or orientations, giving only a partial picture of each material contained in the database.

Genre-specific databases occupy an interesting middle ground. These databases may contain both computational and experimental information, but tend to be focused on a particular type of material, such as structural metals or thermoelectric materials (PRISMS n.d.; Gorai et al. 2016). The TE Design lab is an excellent example of one such database. The TE Design Lab is an online, open-source database that includes both computational and experimental data from thermoelectric materials, along with data visualization and analysis tools. This database permits users to submit new data for review, and provides templates for users to run new calculations generating new data to be included in the database. The inclusion of both experimental and computational data also sidesteps the challenge of small datasets that is common for experimental-only databases. With limited experimental data, predictions are difficult to extend beyond the domain material class, and trends, especially those developed through statistical learning, can be skewed by historically more-studied materials (Gorai Comp Mat Sci 2016). Thus, the TE Design Lab format, which standardizes and curates computational & experimental data inputs and provides users with tools to make data accessible, should be a model for new genre-specific databases in the area of materials for extreme environments.

There are a few examples of extreme-environment materials databases in existence. The high-temperature oxidation database cited in chapter 2 is one example (Barrett 2003). In fact, international (IAEA) and national (EDB) databases exist for tracking embrittlement of structural materials for nuclear reactors, and the U.S. Naval Research Lab has compiled an extensive database of corrosion testing results (IAEA 2009; Wang and Subramani 2008; U.S. Naval Research Laboratory 2018). However, these databases tend to have limited access, with little to no advertisement or description of the data they contain or how it is used. While such barriers may be necessary due to proprietary or security reasons, they hinder collaboration, which limits the value of such hard-won data. Corrosion and radiation especially are challenges that cross the military and civilian communities; and while reactor embrittlement databases do serve a very large

community, they are limited in the data they collect. The EDB does not include microstructural information, and it is unclear to what extent the IAEA reactor embrittlement database does. However, microstructure has a direct effect on embrittlement and avenues of environmental attack (corrosion); therefore microstructural information would be valuable to both of these communities (Allen and Tan 2008). Innovation is required to establish standardized, preferably automated characterization of relevant microstructure for such databases, especially without requiring long-term storage of individual micrographs. Furthermore, databases must provide detailed, well-advertised descriptions of their contents in order to adequately attract a broad and effective user base.

B. Data Repositories

Data repositories are distinct from databases in that they are an intermediary between potential data users and data storage locations. Data repositories provide a way for researchers to make experimental and computational data publicly available and discoverable. Data are often uploaded in a variety of forms, but repositories may also include links to locally hosted datasets, rather than actually storing the data on their own servers. This can minimize the burden to the repository host by decreasing the necessary available data storage infrastructure. Data suppliers are responsible for providing metadata covering the contents of available files and datasets.

One of the most comprehensive materials data discovery sites is the Materials Resource Registry hosted by NIST (NIST n.d.). Beyond linking to materials data sets, it also links to many other materials resources, software, computational tools, and reference databases. Materials data repositories may also be found using the Registry of Research Data Repositories, which is not specific to materials science but does allow searches by subject matter (RE3Data n.d.).

Mat-DB is a database application available to the European Union research community that has the similar goal of aiding data management and discoverability for the materials community (Joint Research Centre n.d.; Over et al. 2005). Mat-DB supports restricted, registered, and open access options for its users, and also promotes standard data-entry formats. Mat-DB is managed by the European Commission Joint Research Center.

Materials Commons is a data repository and e-collaboration platform linked to the DOE-funded PRISMS Center, which focuses on structural alloys. Materials Commons provides ways to track data provenance and scientific workflows, making metadata readily available (PRISMS n.d.). Users can also control access to their projects on the Materials Commons. Although PRISMS began by focusing on structural metals, the tools and protocols developed within the PRISMS framework are envisioned for use for general materials development and discovery. The NIST Materials Resource Registry contains links to both the PRISMS website and Materials Commons.

Unlike NIST and the DOE, the DoD does not have a public site devoted to materials data discoverability and awareness. However, data management plans, including plans to make federally funded, unclassified data publically available, are now a requirement for DoD research

funding proposals (DoD 2015). The Defense Technical Information Center (DTIC) also makes some reports available. It seems impractical for the DoD to establish its own materials resource repository in light of existing open-access options. However, the lack of open-access data centralization in DoD-critical areas such as corrosion, high-temperature, and high strain-rate material testing impedes data discoverability. Therefore, we suggest that program offices specializing in these areas either provide recommendations for particular locations for data storage, or that the DoD build a resource registry similar to the one hosted by NIST that would be searchable by DoD focus areas.

C. Analysis Tools

Many statistical learning and data mining packages are now available for open source programming languages such as R and Python. These include packages for classification, prediction, regression, unsupervised and supervised learning, data preprocessing, and data visualization. Tools specific to materials science include those that focus on analysis methods for data generated by common material characterization methods, including imaging and spectroscopy.

Open source materials data analysis tools can significantly enhance the proliferation of data science methods in this field. Even when access to data is restricted due to security or proprietary reasons, analysis tools are useful and necessary. For example, the PyMKS package mentioned in chapter 2.C can be openly used to quantify microstructural information in a segmented image, or to generate microstructures for further analysis (Brough, Wheeler, and Kalidindi 2017). This tool is specifically aimed at materials science studies, but it can easily be combined with general image analysis tools such as scikit-image and OpenCV, both of which are also open source Python packages. PyMKS is just one tool available from Georgia Tech's MATIN (Materials Innovation Network) project, which also provides links to data repositories and e-collaboration tools (Georgia Tech n.d.).

Another open-source material data analysis platform is DREAM.3D, which allows reconstruction of multi-dimensional data sets, generation of synthetic microstructures, and access to image analysis filters and workflow documentation (Groeber and Jackson 2014). This platform meshes material microstructure so the user can assign descriptors (e.g., grain, dislocation, orientation) to individual mesh constituents.

Open source tools like those available from DREAM.3D for microstructure characterization of multi-dimensional imaging data are sorely needed. Reliable tools for recognition of dislocations, grain boundaries, precipitates, pores, or other features are essential. Commercial software exists for conducting such image analysis, but the lack of open-source tools stunts materials research by limiting access and collaboration. Being able to analyze more images enhances the value of a single, often expensive microscopy session, and increases the amount of data a single research can gather and analyze in a fixed time. A single researcher may collect dozens of images during a typical two-hour electron microscopy session—but rarely analyzes every image in detail, thereby

limiting the statistical value of the dataset. User facilities such as National Labs are ideal places to develop such tools. The Center for Data-Driven Discovery at Brookhaven National Laboratory focuses on challenges related to National Laboratory data generation and analysis, including for transmission electron microscopy analysis and X-ray scattering (Brookhaven National Laboratory n.d.). Like image analysis, spectral analysis, such as curve-fitting and peak identification, applies to many types of materials characterization data.

Publicizing the availability of analysis tools, databases, and data repositories will accelerate their use by the materials community. Materials education programs in particular are well-positioned to inform students and staff about the existence and application of such resources. This in turn will encourage students to take advantage of these tools in their own work, enhancing the materials workforce skill base, promoting standardized data analysis methods and reporting, and promoting novel tool development.

D. Perspectives from Test Labs and Experts

Government laboratories have been characterizing materials for decades. Testing may be motivated by different reasons, such as understanding and predicting material failure in government property, searching for new and improved materials, or evaluating potential government acquisitions. The results of such testing may be classified, and therefore not available to the public. However, this does not preclude us from considering the best practices for data storage, curation, maintenance, and sharing within DoD infrastructure. Given the history of such testing and characterization, it is likely that considerable knowledge has been lost through lack of digitization in a searchable format. This section summarizes conclusions based on feedback from a few individuals involved in government material testing and characterization. *However, we recommend a full survey of DoD and/or government materials testing & characterization data management protocols in order to identify best practices and gaps in this area*.

Common practices for data storage, maintenance, and sharing within government material testing communities were identified based on feedback from individual members of four government material testing groups.[2] For the most part, data are collected and stored by individual teams following whatever format the team deems appropriate. Often, the results of data analysis, along with the most relevant supporting data, will be published in either open literature or technical reports. This publication may require approval by the funding sponsor, depending on the specific project. The quantity of data reported, and whether data are reported as summaries or as tables of individual observations, is highly dependent on the author and the purpose of the report. For example, when conducting quality assurance testing, one might report summary statistics (mean and standard deviation) rather than individual test results. When reporting on the progress of a

[2] Dr. James Smialek, Material High Temperature Oxidation Group, NASA Glenn Research Center; Dr. Farrel J. Martin, Code 6134, Center for Corrosion Science and Engineering, Naval Research Laboratory; Jay Ong, Corrosion testing, NSWC Carderock; Joel McDuffee, Nuclear Experiments and Irradiation Testing Group, Oak Ridge National Laboratory

long-term testing program, one might report tables of measurements. The DoD has made efforts at materials data longevity and accessibility in supporting Military Handbooks and Information Analysis Centers, but these do not facilitate straightforward access to the kinds of data sets most useful for statistical learning efforts (Information International Associates 2012).

In some cases, researchers have compiled data collected over multiple years in the form of a personal spreadsheet or database. The oxidation data from NASA is one example of such an effort, comprising decades of material testing data stored in a standardized data format that was made available to the high temperature oxidation team (Barrett 2003). Such compilations are useful to the teams that continue testing or conduct future analysis, but can present limitations for data sharing between researchers or groups that use different formats or slightly different testing protocols. This is especially true when there are access restrictions on the data in question; in such a case, a formal collaboration or release request may need to be established before data can be shared. Groups with such internal protocols may find it worthwhile to publicize the data storage format (without releasing actual data) to lay the groundwork for potential standardization in the event of future collaborations or data sharing.

Beyond their value as accessible data sources, internally-maintained databases can encourage institutional protocols that preserve data for long term effectiveness across multiple testing efforts. For example, the Barrett oxidation database has inspired a similar development in the area of environmental barrier coatings within the same laboratory at NASA. We note, however, that significant barriers exist in DoD labs to implementing internal databases that go beyond a simple spreadsheet format. Cited challenges include lack of software, data storage architecture, established data storage formats, or incentives to conduct data entry on significantly backlogged data. The lack of an actively maintained database can lead to data loss as storage formats and operating systems evolve over time and personnel turn over. One way to prevent such data loss is to implement institutional guidelines that encourage data preservation actions, especially in organizations where the loss of individuals can mean the loss of decades' worth of data. For example, when possible and approved for public release, data may be preserved by uploading onto a data repository similar to the ones described in chapter 3B. Testing data for materials in extreme environments are particularly worth preserving, because the testing is often expensive, time-consuming, and highly technical.

A second way to enhance the value of such datasets is to use e-collaboration tools to facilitate collaboration within government agencies or between government labs. Access-controlled tools like those available through PRISMS and MATIN can help users develop customized data entry protocols that allow compilation of data from different testing labs with similar interests. Similarly, automated analysis codes produced using Python, DREAM.3D, or other platforms can be exported between labs to ensure consistent characterization protocols.

There is also room for improvement when moving forward from historical datasets. New datasets need not be restricted to only the data collected in the past. For example, microstructural information went largely unreported for the oxidation data set and micrographs were not included

in the database (Barrett 2003). However, information such as phase fractions or phase length scales could be quite useful, for example to determine the limits of potential improvements in oxidation resistance, or to identify helpful and harmful scales. Along this vein, in the absence of experimental data, the results of computational simulations could be used to augment the existing database.[3] This is another way in which historic data can remain relevant—by providing straightforward validation and concrete targets for ongoing simulation efforts.

Although enterprising researchers have occasionally been able to ensure the persistence of their data through decades, such efforts would be more successful and widespread with increased availability of data management tools and institutional support for data preservation. Experimental data retain value over time as new analysis and simulation methods are developed, but only if they remain accessible and documented. Data preservation also guards against unnecessary duplication of effort. Therefore, organizations that conduct extensive materials testing in extreme environments will benefit from supporting their analysts and researchers in efforts to preserve expensive, hard-won data. Supporting actions might include: (1) providing guidance on methods and means of data storage and discoverability, (2) contracting for development of internal data repositories and e-collaboration tools, (3) supporting committees to establish standardized data entry formats and protocols within each material testing community, and (4) incorporating data management plans into proposals for new research or testing programs.

[3] Based on discussion with Dr. James Smialek, NASA

4. Target Areas

In this section, we describe the factors that contribute to a materials data science problem, and highlight a few target problems in the area of materials in extreme environments that appear well-suited to a data science approach.

A. Structuring a Materials Data Science Problem

When structuring a study using data science tools, the following list of questions should help guide choices related to data gathering, analysis methods, and variable selection. These questions need not be asked in the order listed, as analysts may not have control over data collection.

1. Is the study explanatory, utilitarian, or exploratory?

An explanatory study seeks to understand which knobs to turn to control a particular outcome, and to explain why those knobs might work. In contrast, a utilitarian study may aim to develop a model that will predict an outcome with available information, without necessarily explaining connections. Finally, an exploratory study may not have a defined goal, but may seek to identify previously unknown relationships or structure within a data set. A research study will often be explanatory or exploratory in nature, while process control development may be more utilitarian. Knowing the answer to this question will help the analyst select appropriate statistical learning tools.

2. What is the material domain space?

This will likely be geared by the application (i.e., corrosion-resistant alloys, high temperature structural materials). Highly disparate mechanisms may be at play within different material classes (e.g., metal vs. polymer) or forms (e.g., single crystal vs. polycrystalline). It may be the case that multiple material classes are of interest to solve a problem (e.g., lightweight structural materials can include composites or aluminum alloys), but the disparity in synthesis mechanisms and microstructure would suggest using separate statistical learning modules for each material class. The material domain may be extremely narrow, as for optimizing a processing line, or extremely broad, as for designing high entropy alloys. It is important to understand the breadth of potential materials of interest, and what sources of variation may come into play. The answers to this and the following question will also be important when defining data entry formats and options for new databases.

3. What are the properties of interest?

Perhaps we want to predict a property based on a known composition or microstructure. Perhaps we want to describe a microstructure given an image, with an eye to a particular application or theoretical model. Perhaps our goal is to understand how sensitive a particular property is to variations in microstructure or processing. We may want to conduct unsupervised learning on a suite of data with unknown value. Whatever our goal, we need an idea of which variables are going to be considered in our data analysis, and of what (if anything) we are trying to optimize. These may include mechanical or transport properties, composition, microstructure, cost, processing parameters, or environmental parameters. If the problem at hand is linked to a particular application, this will define the suite of properties that matter. It may be necessary to define a metric or score to describe how well a material satisfies some set of requirements.

4. Where will data come from, and what data quality, contents, and quantity are expected?

Whether data are available from the literature, are contained in an existing database, are to be gathered from experiments, or are to be generated via computational means, it is important to have an understanding of the expected number of data points, level of accuracy or trust in those data points, and the actual types of data available (e.g., images, measured properties, processing histories). Subscription or registration-based databases do not always publicize a detailed description of their contents, which creates a barrier to researchers who are deciding whether to gain access to those databases. The expected number of data points (n) and number of descriptor variables (p) will also inform the choice of learning method.

B. Specific High-value Targets

This section will highlight several materials in extreme environments problems of interest to the DoD that seem well-suited to a data-science approach. For each target area, we address to the extent possible the questions posed in chapter 4A, including articulating key explanatory variables and target properties.

1. Complex Concentrated Alloys (CCAs) and High Entropy Alloys (HEAs)

CCAs and HEAs are metallic alloys that incorporate a high degree of disorder and multiple base elements in their microstructure. These alloys are of interest for numerous applications within the DoD. (Miracle 2015, 2018; Zhang et al. 2014) HEAs, typically defined as containing at least 5 principal elements at concentrations between 5 and 35 atomic percent, are a subset of CCAs. While HEAs are often described as single-phase solid solutions, CCAs may incorporate multiple phases or have fewer than five principal elements (Miracle 2018).

Relative to traditional alloys made by adding small amounts of impurities to a one- or two-element base metal, CCAs and HEAs constitute a combinatorial explosion in terms of potential compositions. Because of the enormous design space, including through adjustment of processing

approaches, CCAs are being considered to solve myriad extreme environment challenges that confound existing alloys and superalloys, including high temperatures, low temperatures, corrosion, radiation, and wear (Zhang et al. 2014). The opportunities for data science solutions are likewise extensive. Each part of the PSP linkages needs to be established for these alloys, from predicting fundamental crystal structure based on composition (see chapter 5A) to predicting properties of interest or microstructure based on composition and processing parameters, to identifying the relationships between microstructure and target properties. There is also ongoing work in simply establishing models for transport and deformation mechanisms in CCAs. Although many such models exist and are well understood for conventional alloys, they may not translate to highly disordered systems. Without hard models, soft models based on statistical learning from high-throughput computational and experimental data may point the way to greater understanding.

For HEAs and CCAs, research is still at an early stage, which means that explanatory and exploratory models will be most useful. Particular focus groups will be determined by the application—for example, alloys with BCC crystal structure are preferred over FCC alloys for structural applications that require higher strength and lower ductility (Zhang et al. 2014). Variables of interest for optimization may include mechanical properties such as weight-normalized yield strength or stiffness, maximum use temperature, diffusion coefficients, or corrosion resistance. Explanatory variables for computation-based phase prediction should be those that can be determined from composition alone, such as average atomic size differences or valence electron concentration (Guo 2015; Yang and Zhang 2012). For microstructural prediction, processing parameters may be of interest, such as heat treatment information (time, temperature, cooling and heating rates) or processing method (e.g., arc melting, casting, mechanical alloying). Microstructural variables could include grain size, phase fractions or length scales, porosity, precipitate morphology, or dislocation density.

Some work has been done generating thermodynamic databases for high entropy alloys by commercial enterprises such as Thermocalc and Ques-Tek (Mao, Chen, and Chen 2017; QuesTek Innovations LLC 2018). However, experimental data, particularly with regard to microstructure, is needed to produce a truly effective data infrastructure for HEAs and CCAs. High-throughput synthesis and characterization is challenging, especially for structural materials, where many property measurements are needed and test miniaturization is restricted by microstructural length scales (Miracle 2015). This is a case where common data reporting by multiple labs could grow a knowledge base that would otherwise be very difficult to compile. Because this research field is still young and the research community is relatively small, there is real potential to develop common data-reporting schemes now (perhaps associated with publication) that could truly benefit the entire community. HEAs could represent a flagship area for encouraging data-reporting habits common in other fields to transition to materials scientists.

2. Corrosion-resistant Materials

Corrosion can be broadly defined as material degradation in response to chemical attack (M&MP COI 2017). Corrosion causes material loss and transformation (often, oxidation), change in material properties, and can participate in crack growth when coupled to mechanical stress. All of these factors amount to damage that weakens structures and leads to eventual material failure. Although we are most often concerned about corrosion in metals, it can also be a problem for ceramics such as SiC (Allen and Tan 2008). The exact flavor of corrosion will depend on environmental factors like pH, the presence of sulfur, hydrogen, or oxygen, and additional factors such as high temperatures or mechanical stress. Furthermore, microstructural variables such as grain boundary character, grain size, and residual strain can directly influence the mode and rate of corrosion (Allen and Tan 2008). With the diversity of potentially corrosive environments and corroding materials of relevance to DoD infrastructure, corrosion prevention and control is a complex, ongoing challenge (PEO Land Systems 2017). Thus, corrosion is a cross-cutting problem for the DoD, constituting its own Technical Area Team within the Materials and Manufacturing Processes Community of Interest (M&MP COI 2017). Therefore, the design and testing of corrosion-resistant materials, along with prediction of corrosion in operating environments are ongoing goals for DoD labs (M&MP COI 2017).

As an example of the kind of corrosion problem that may benefit from a statistical learning approach, we consider zirconium alloys for nuclear fuel cladding in fission reactors. Such cladding undergoes corrosion in both light and boiling water reactors, and it has been observed that both compositional and microstructural changes can significantly affect the kinetics of corrosion and oxide growth (Motta, Couet, and Comstock 2015). Alloy degradation may be described by a number of features, including hydrogen pickup fraction, pre-transition times, post-transition corrosion rates, and pore formation and growth rates. The transition is a feature of oxide growth kinetics that demarcates a change from parabolic or subparabolic corrosion kinetics to much faster scale (oxide) growth. Often, such transitions are periodic and oxide growth is a cyclical process. Alloys with longer pre-transition times and slower scale growth are considered more corrosion-resistant.

The orientation and anisotropy of growing oxide grains, along with pore growth, contributes to stress accumulation that can affect transition time and thickness. However, the connection between the alloy composition and microstructure and the oxide microstructure is not well understood (Motta, Couet, and Comstock 2015). Nor is the role of precipitates (secondary phases) in the alloy fully accounted for. Precipitates may oxidize at delayed rates as defined by Pourbaix or Ellingham diagrams. Nonetheless, Zr alloys with enhanced corrosion resistance tend to have both complex composition and secondary phase precipitates. Therefore, variables of interest for modeling corrosion kinetics might include microstructural descriptors of precipitate shape, size, or distribution, in addition to typical descriptors like grain boundary density. Compositional variables should also be used as descriptors, along with exposure variables (e.g., temperature, time, hydrogen levels, applied stress).

Prediction of kinetic parameters (e.g., coefficient and exponent for pre-transition phase) or classification of oxide growth kinetics (immediate breakaway, single cycle to breakaway, or fully cyclic) are both potential goals for statistical models from Zr alloy corrosion data. A database compiled from documented micrographs, compositions, and oxide growth kinetics could be analyzed to attempt to model some of these effects and guide further alloy development and mechanistic hypothesis testing.

For structural materials of relevance to the DoD, material testing data gathered by members of the Naval Research Lab or testing and certification groups at the Naval Surface Warfare Centers could similarly be analyzed with a statistical learning approach in order to explain or explore the mechanisms of corrosion in operational environments. In addition, utilitarian models based on machine learning from naval corrosion testing databases may be useful for flagging or predicting required maintenance checks for in-service systems. Some work to this effect has been done to predict corrosion rates of common alloys based on worldwide measurements of corrosion rates and weather variables at military bases (Morefield et al. 2009). Efforts like these should continue and be expanded to newly introduced alloys and coating systems, particularly with reference to initial compositions and microstructures. Furthermore, similar efforts could be applied to building models to predict high temperature oxidation, which is an important problem for turbines in jet engines (M&MP COI 2017; Smialek et al. 2000).

3. Thermal Barrier Coatings and High Temperature Structural Materials

Hypersonic vehicles face a major challenge with high temperature resilience (Sziroczak and Smith 2016). Atmospheric drag during ascent or cruise can generate significant heat that is proportional to the atmospheric density and the cube of velocity. While only about half of this heat actually reaches the aircraft, it means that materials are needed that can withstand very high temperatures, with the maximum usage temperature directly affecting the possible speed of the vehicle at various altitudes. To address this challenge, hypersonic vehicles require high temperature structural materials and/or thermal protection systems (TPS) that dissipate heat, including by ablation. The leading edges of hypersonic vehicles experience the most intense heat.

Materials of interest for these purposes include C-C composites (for both leading edges and surfaces, maximum T ~1500 ºC), refractory metals (e.g., Inconel, maximum T ~1400 ºC), and ultra-high temperature ceramics (usually only for leading edges, e.g., Hf/Zr diboride, maximum T ~ 1200-1600 ºC) (Sziroczak and Smith 2016). For leading edge materials, thermal diffusivity and melting point should be maximized. For structural materials, mechanical properties at high temperatures are also important, such as interlaminar toughness (for layered composites), shear strength, and tensile strength.[4]

Explanatory variables will depend on the material class. For example, variables such as pore density, fiber diameter, interphase coatings, fiber tow properties, fiber weave, and fiber alignment

[4] Based on conversation with Dr. William Hong, IDA

are meaningful for C-C or metal-matrix fiber composites, while variables such as precipitate shape and dispersion may be important for oxide dispersion strengthened Ni alloys. Inconel is in fact a family of Ni-Cr alloys including varied composition and microstructure.

Utilitarian models that can relate processing conditions to microstructure (e.g., pore size and distribution) and microstructure to mechanical and thermal performance could be quite valuable. Explanatory models connecting composite, ceramic, and metallic microstructure to thermal diffusivity would likewise be of interest. However, data for hypersonic materials is scarce in the public domain. Nonetheless, this is an area where statistical learning could be applied to either optimize processing or assist with material design.

4. Radiation-resistant Structural Materials

While not as widespread of a challenge as corrosion, irradiation is also of relevance to the DoD, particularly for structural materials in reactors onboard nuclear-powered submarines and ships. Like corrosion damage, radiation damage is controlled by complex processes that connect material composition, microstructure, and exposure conditions. In fact, structural materials typically have operating temperature *windows* that define both a maximum and a minimum usage temperature. This is because the mechanisms of radiation damage and resistance differ depending on temperature; these processes include defect formation and annihilation, diffusion, absorption into defect sinks, accumulation, and amorphization. Radiation resistant structural material classes include ferritic steels, V alloys, and SiC/SiC composites. The strategies to design radiation resistance tend to focus on adding radiation-resistant matrix phases, immobilizing point defects, and engineering high-sink-strength microstructures. Sinks may be cavities, precipitates, dispersoids, grain boundaries, or dislocations (Zinkle and Snead 2014).

It is easy to see how valuable consistent, quantitative microstructural descriptors would be in this context. For radiation damage, the average spacing between defects can be just as important as the size of those defects, because a balance must be struck between mechanical properties (e.g., toughness, strength) and radiation tolerance (e.g., operating temperature range, ductile-to-brittle transition temperature as a function of dosage). Material healing can be governed by how readily a defect can migrate to a sink. It is difficult to control microstructural variables independently, even at fixed composition. With enough data points, a wide microstructure space can be covered and trends identified, even in the absence of systematic variable adjustment.

Data on radiation embrittlement has been gathered for decades in reactor pressure vessel (RPV) surveillance databases (Wang 2010; IAEA 2009). Such data mean that the effects of *exposure conditions* and to some extent *composition* on RPV steels are well understood. Microstructural data, or at least micrographs, may also exist from efforts like the UCSB Radiation Embrittlement Damage Analysis and Predictions (REDAP) project. Research efforts to understand radiation damage in particular material classes continue, for example at Oak Ridge National Lab. A database of micrographs could be analyzed with computational image processing to quantify variables like the concentration, morphology, and distribution of features like pores (and voids),

dislocations, grain boundaries, or precipitates. Paired with starting compositions of samples, surveillance data on exposure conditions and mechanical property measurements could be used to build predictive models or explanatory models that relate embrittlement to particular microstructural characteristics. Such analysis could also help link the propagation of variance between processing, microstructure, and performance, which could inform questions about the *importance* of certain variables even if it does not explain the *mechanisms* of importance.

5. General Trends

A statistical learning approach is best aimed at materials problems where physical mechanisms are poorly understood or difficult to model with hard modeling. In such cases, soft modeling may significantly reduce computational effort, or may point to physical effects that were previously difficult to resolve due to conflicting or correlated variables (Rajan 2015). Problems in which the microstructure acts as an important variable fall squarely in this category, because complex microstructures are extremely difficult to model. Microstructure is an essential component of design for corrosion and radiation-resistant materials, and it plays an important part in defining thermal diffusivity and mechanical properties. Likewise, when manufacturing materials with complicated microstructures, one must understand what makes two samples the same or different, and one must understand what level of variance actually matters for material performance. The more complicated the material, the more difficult or time-consuming it is to quantitatively characterize the microstructure—which advocates for including automated image processing into any quality assurance workflow.

This chapter has focused on the role of *experimental data*, but it should be noted that *computational data* does have a role to play for materials in extreme environments. In particular, simulating microstructures (to get around the problem of independent property control) can be a powerful technique for generating data on relationships between microstructural variables and target properties. In such a case, one needs to have a physical model in mind to determine a target property (e.g., tensile strength) from a known microstructure, but advances in finite element modeling and Monte Carlo simulation have made it possible to do this in some cases where a deterministic relationship is uncertain or inaccessible analytically (Niezgoda, Kanjarla, and Kalidindi 2013). This approach is also worthwhile when trying to determine how a particular mechanistic hypothesis might translate to observable effects.

The high entropy alloy problem represents a case where there is a dearth of physical knowledge and an enormous space of possible alloys. In fact, similar concepts are also being extended to ceramics, with numerous potential applications for materials in extreme environments (Gild et al. 2016; Rost, Christina M. et al. 2015; Zhou, Jieyang et al. 2018). Because we have not determined the limits of these materials, we do not know what they are capable of, and they are therefore being investigated for nearly every extreme-environment application. The sheer lack of knowledge means that data science methods may be able to help narrow down the search space in

various application areas. However, here and for any analysis built on experimental data, it will be important to guard against bias resulting from a limited domain space of experimental results.

5. Case Studies

A. High Entropy Alloy Classification

In this chapter, we apply a few classification methods to experimental data on cast high entropy alloys. The data come from a table available in previous work that sought to classify alloys by their ability to form solid solutions based solely on composition (Yang and Zhang 2012). That study found that the average atomic size difference δ and the ratio of entropy and enthalpy of mixing Ω could be used to discriminate between solid-solution and mixed-phase compositions. Here, we focus on discriminating between particular phases—face-centered cubic (FCC), body-centered-cubic (BCC), and intermetallic phases labeled by α and σ.

Phase selection rules for HEAs have been developed in a number of studies (Guo 2015). For example, Dominguez et al. showed using principal component analysis that enthalpy of mixing and valence electron concentration can be helpful predictors of BCC or FCC character (Domínguez, Goodall, and Todd 2015). That study also included entropy of mixing, average atomic size difference, and electronegativity difference as predictor variables. In slightly different work, high throughput computations of enthalpies of formation of binary phases were used to predict combinations of elements that could form single phase HEAs (Troparevsky et al. 2015).

The strategy in the present work is to explore the strengths and weaknesses of a few multi-label classification schemes using data points containing composition and phase information for 110 cast high entropy alloys. The 110 data points come from Table 1 of reference (Yang and Zhang 2012), after excluding those data points that contained phases other than the four of interest (FCC, BCC, α and σ). Observations labeled as BCC1 + BCC2 (originally distinguished by containing BCC solid solutions with fluctuations in local lattice parameter and composition) were counted as BCC. Other phases were excluded due to limited occurrence in the data set (< 4 examples). Rather than focusing on specific elemental content, we focused on descriptor variables that could be determined for arbitrary composition, following (Yang and Zhang 2012) and (Guo 2015). The chosen descriptor variables included those highlighted in previous studies and a few additional variables; particularly *J*, which is the average total angular momentum quantum number based on the composition. Table 1 reports the descriptor variables used for this analysis, including the formula for calculating these given a composition. Note that the variable "Precursor" is categorical with up to three possible values—this translated to three dummy variables, which brought the total number of variables up to 12.

We used a multi-label classification approach. As highlighted in chapter 2A, this approach has non-exclusive class labels built into its framework, rather than defining exclusive classes with some uncertainty about class membership. For the present data set, this means that any observation

can have up to four labels, rather than defining separate classes based on (possibly sparse) combinations. For example, we do not need separate classes for FCC + σ, BCC + σ, FCC + BCC, and FCC + BCC + σ. The multi-label approach is intuitive to this problem, because it is known that samples can contain multiple phases, and transitions from one phase to another will often be systematic (i.e., proceeding from one phase to another via a mixture of phases).

SVMs, decision trees, and random forests were chosen as potential classification methods. Prior to conducting analysis, the 110 sample data set was randomly divided into training and test sets, leaving 93 data points for training and 17 for test. The test data were set aside for final classifier performance analysis. All analysis for this work was done in Python.

Table 1. Descriptor variables for high entropy alloy study

Variable	Units	Description	Formula (if applicable)
δ	%	Average atomic deviation	$\delta = \sqrt{\sum_{i=1}^{n} c_i \left(1 - \frac{r_i}{\bar{r}}\right)^2}$
ΔH_{mix}	kJ/mol	Enthalpy of mixing	$\Delta H_{mix} = \sum_{i=1, i \neq j}^{n} 4 \Delta H_{ij}^{mix} c_i c_j$
ΔS_{mix}	J/(mol*K)	Entropy of mixing	$\Delta S_{mix} = -R \sum_{i=1}^{n} c_i \ln(c_i)$
T_m	K	Melting point	Rule of mixtures
Ω	unitless	Solid-solution formation parameter	$\Omega = \frac{T_m \Delta S_{mix}}{\lvert \Delta H_{mix} \rvert}$
Alloyn	count	Number of constituent elements	Count from formula unit
VEC	count	Average valence electron concentration	Weighted average of element VEC by atomic fraction
J	unitless	Average total angular momentum quantum number	Weighted average of element *J* by atomic fraction.
dEN (Δχ)	unitless	Electronegativity deviation parameter (Pauling scale)	$\Delta \chi = \sqrt{\sum_{i=1}^{n} c_i (\chi_i - \bar{\chi})^2}$
Precursor	categorical	Dominant structure type of base elements	BCC, FCC, or BCC_FCC if equal concentrations

*In the above table, ci and ri denote atomic fraction and radius of element i. R is the gas constant. The first five parameters were available from Table 1 in reference (Yang and Zhang 2012). Electronegativities of elements are based on the Pauling scale. dEN was determined using the same formula as reference (Guo et al. 2011).

Prior to SVM analysis, data were scaled so that each of the 12 variables had mean 0 and standard deviation 1. Following this, principal component analysis was used as a dimensionality-reduction method. While the SVM approach is well-designed to manage multi-label data, it suffers in computational speed with additional variables. Figure 1 plots explained variance against the

number of principal components. Over 95% of the variance in the training data was explained with 7 principal components, so 7 components were selected moving into SVM analysis.

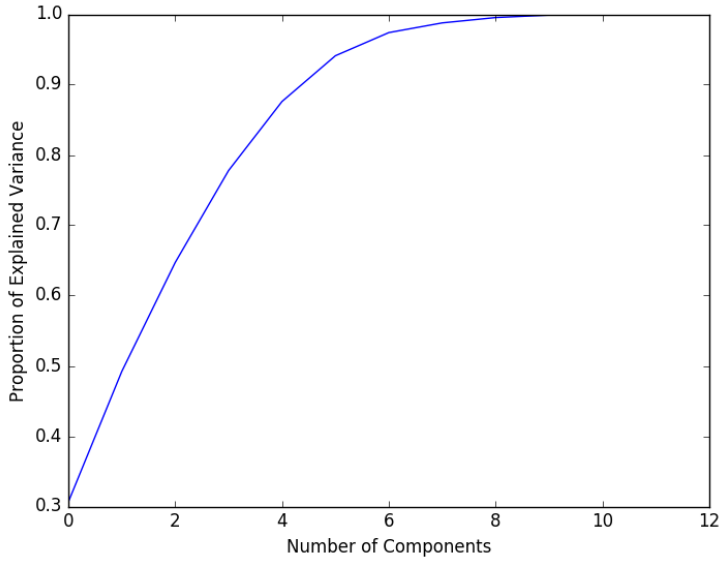


Figure 1. Explained variance vs. number of principal components for the HEA training data

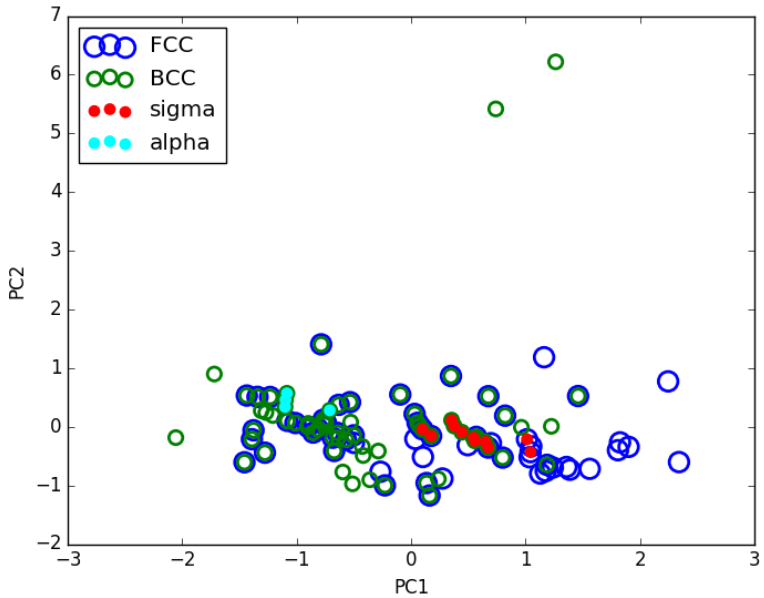


Figure 2. HEA training data plotted against the first two principal components

Figure 2 shows the training data plotted against the first two principal components. The four labels clearly overlap—there are many examples of simultaneous FCC and BCC, and no examples of σ or α phases alone. Furthermore, the α phase only constitutes four observations. We note that because of the sparseness of α data, it did not appear in the test data set. However, we did use 10-

fold cross-validation to build all classifiers in order to give some information about the ability of the classifiers to detect the α phase.

SVMs were trained on each class in a one-vs-rest procedure using either the conventional linear boundary or the radial basis function (RBF) kernel. These kernels effectively adjust the shape of the boundaries that will be produced by the SVM. In the SVM algorithm, a complexity parameter *C* determines how complex the boundary can be. In fact, the larger the value of *C*, the more observations are penalized for being on the wrong side of the boundary. This parameter is determined by cross-validation, and essentially guards against over-fitting the data. Figure 3 shows the cross-validation performance of the SVM classifier versus the value of C for the two kernels. The RBF kernel shows higher accuracy than and similar variance to the linear kernel. Based on cross-validation, $C = 1000$ and 100 for the RBF and linear kernels, respectively.

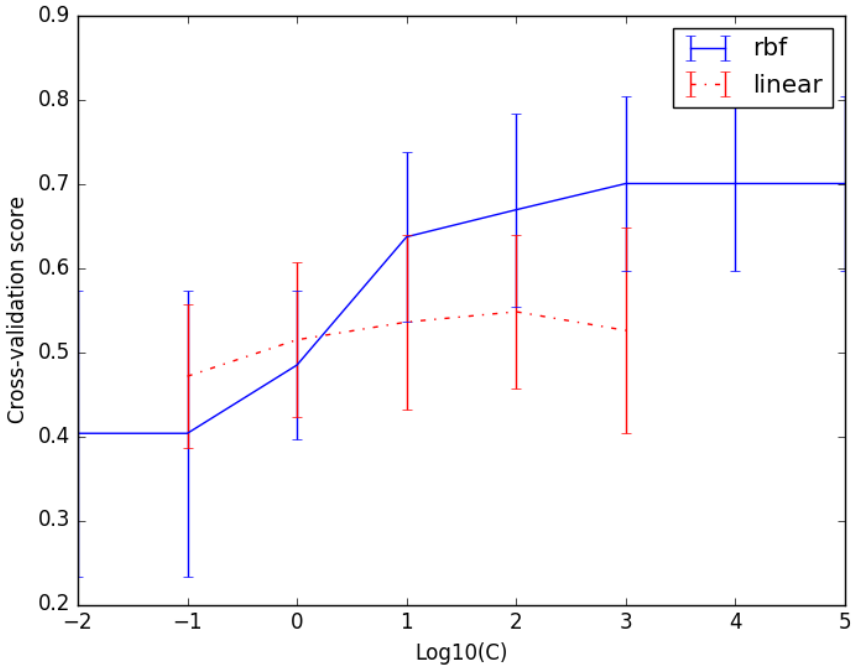


Figure 3. Cross-validation scores vs. complexity parameter C for support vector machine classifiers using the linear or radial basis function (rbf) kernel

The cross-validation results show that the RBF SVM should have accuracy of 60-80% when applied to test data, while the linear SVM should have about 45-65% accuracy. For multi-label classification, accuracy describes the number of points for which class labeling was perfect. In other words, observations that miss one label (false negative), include one incorrect label (false positive), or have two or more of these errors would all be equally counted as inaccurate. Multi-label classification allows for other definitions of performance, however, such as base-class precision (proportion of predicted instances of a label that are correct), base-class recall (proportion of true instances of a label that are correctly classified), and alpha-evaluation. Alpha-

evaluation is a metric that allows one to disparately weight errors of different types (false positives, false negatives) (Boutell et al. 2004).

Table 2. Precision and recall of SVM classifiers applied to HEA data

Class	Precision (RBF)	Recall (RBF)	Precision (Linear)	Recall (Linear)
FCC	*0.87 ± 0.11*	*0.87 ± 0.17*	0.83 ± 0.12	0.78 ± 0.14
BCC	0.93 ± 0.08	0.89 ± 0.14	*0.95 ± 0.08*	*0.91 ± 0.13*
σ	*0.75*	*0.875*	*0.75*	*0.875*
α	*1 false +*	*0.75*	2 false +	*0.75*

*Italics denote the best performance for each metric and class

Table 2 compares the precision and recall of the linear and RBF SVMs in cross validation. The main difference between the two kernels is that the linear classifier had better precision and recall for the BCC phase, while the RBF classifier did better for the FCC phase. The margin of difference between the two models is larger for the FCC phase (about 4–9%) than for the BCC phase (2%). The RBF classifier also did slightly better with avoiding false-positives for the α phase, which only appeared in four of the 10 cross-validation folds. Overall, the RBF classifier has the best accuracy and the best performance across the metrics for the two SVM models.

Decision trees were explored separately for this analysis. Decision trees do not require that data be scaled prior to analysis. To preserve the descriptive nature of the decision tree (a strength of this type of model), dimensionality reduction using PCA was also not conducted. Decision trees were grown using the Gini impurity criterion to determine the positions of category splits. The tree depth and minimum number of samples per leaf was tuned by cross-validation, as shown in Figure 4. Based on this analysis, there were two possible *best* trees: one with depth 6 and at least 3 samples per leaf, and one with depth 9 and at least 1 sample per leaf. Of these, the first is a simpler model but the second shows slightly smaller variance in prediction accuracy.

Table 3 compares the cross-validation precision and recall of the two types of tree. The 9-1 tree performs best or ties for best performance in all metrics, with a general better precision than the 6-3 tree. In particular, the 9-1 tree does a better job correctly labelling the σ phase, which likely follows from its ability to filter to finer individual leaves.

The final 9-1 decision tree trained on the full training set has its first split for δ at 5.04, and the next level of splits for *J* (2.878) and melting point (1778 K).

Table 3. Precision and recall

Class	Precision (6-3)	Recall (6-3)	Precision (9-1)	Recall (9-1)
FCC	0.80 ± 0.13	*0.84 ± 0.16*	*0.82 ± 0.17*	*0.84 ± 0.16*
BCC	0.90 ± 0.09	*0.94 ± 0.07*	*0.92 ± 0.09*	*0.94 ± 0.07*
σ	0.63	0.75	*0.67*	*0.94*
α	1 false +	*0.25*	*0 false +*	*0.25*

*Italics denote the best performance for each metric

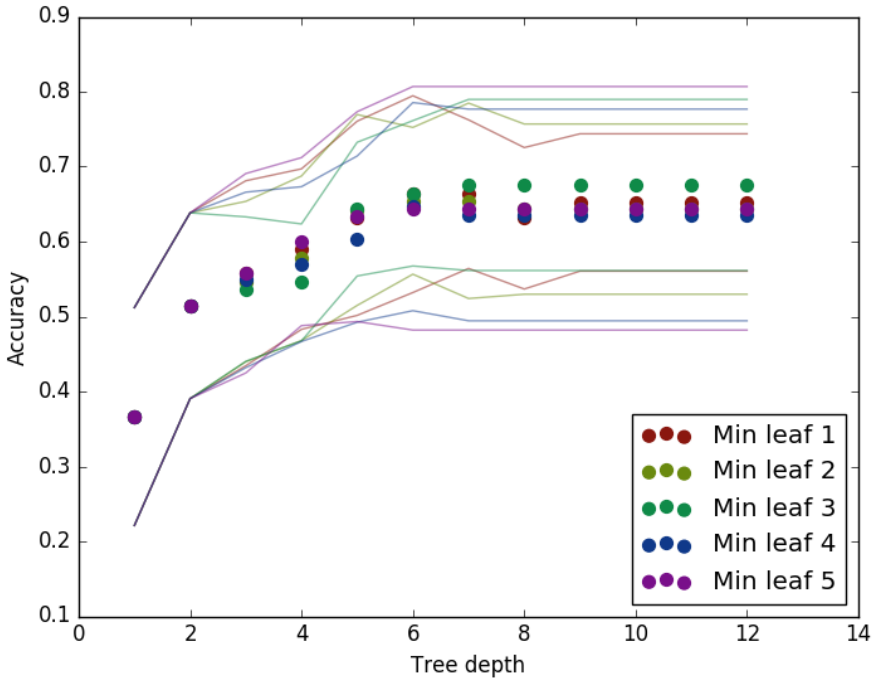


Figure 4. Cross-validated accuracy of decision trees as a function of tree depth and minimum number of samples per leaf (Min leaf #)

A random forest algorithm was also applied to this data set. The random forest algorithm builds an ensemble classifier by repeatedly fitting a weak classifier to a data set that has been randomly resampled with replacement. For the present study, cross-validation was used to determine the optimal weak classifier depth of 6. The random forest classifier has a cross-validation accuracy of 65–82%, which is a reduced variance as compared to a simple decision tree classifier. A plot of variable importance for this classifier identifies the four most important variables in order of decreasing importance as follows: *J*, VEC, Ω, and δ. This result fits well with intuition from Yang et al. Note that *J* and δ were two of the top three variables in the 9-1 decision tree. The random forest variable importance accounts for appearances like this over many iterations of training similar classifiers. It is interesting to note that the average angular momentum quantum number *J* was not included in the referenced previous work classifying HEA crystal structure, but is one of the most important predictors for the tree and forest models.

Table 4 and Table 5 compare the precision and recall, respectively, of the SVM-RBF model, the 9-1 decision tree, and the random forest model, determined by 10-fold cross validation. Because the α and σ phase did not occur in the test sets for all 10 folds, recall is determined based only on the folds in which these phases appear, and the false positive rate (number of false positives in all folds) is reported instead of precision for the alpha phase. The best-performing score of the three models for each class is shown in italics.

Table 4. Precision of classification applied to HEA data

Class	SVM-RBF	9-1 decision tree	Random forest
FCC	*0.87 ± 0.11*	0.82 ± 0.17	0.83 ± 0.11
BCC	*0.93 ± 0.08*	0.92 ± 0.09	0.92 ± 0.09
Sigma	0.75	0.67	*1*
Alpha – false positive rate	1	*0*	*0*

* Italics indicate the best score for the row

Table 5. Recall of classification applied to HEA data

Class	SVM-RBF	9-1 decision tree	Random forest
FCC	0.87 ± 0.17	0.84 ± 0.16	*0.92 ± 0.10*
BCC	0.89 ± 0.14	0.94 ± 0.07	*0.97 ± 0.06*
Sigma	0.875	*0.94*	0.875
Alpha	*0.75*	0.25	*0.75*

* Italics indicate the best score for the row

The tables indicate the SVM-RBF model has the best precision for the FCC and BCC phases, with moderate precision for the alpha and sigma phases. The random forest model has the best recall for all but the sigma phase, and is tied with the SVM for the sigma and alpha phases. The random forest model also exhibits the smallest variance in recall values. Although the random forest model sacrifices some precision for the FCC phase relative to the SVM model, it has the best performance across all metrics. Therefore, if one model were to be selected as a future classifier, we would choose the random forest model.

As a final confirmation of this model selection, we retrained these three models on the full training set and applied them to the test data. All three models had the same prediction accuracy (65%) and perfect recall for the BCC phase, with the most common error being incorrect classification of FCC or BCC as both FCC and BCC. Both the RBF SVM and the decision tree had 91% recall of the FCC phase and accurately captured the one instance of σ phase in the test set. Compared to the RBF SVM, the 9-1 tree was more even-handed about its false-positives for the FCC and BCC phases (FCC and BCC precision of 83 and 81%, respectively, for the tree vs. 77 and 87% for the SVM). In contrast, the random forest had perfect recall for the FCC phase, with similar but slightly better precision than the SVM (79 and 87%), and missed the instance of

the σ phase in the test set. Improvements to these models might be possible by changing the scoring function during model training and cross-validation, so that false positives were more heavily penalized than false negatives.

Overall, this case study showed that support vector machines, decision trees, and random forests could be used to construct adequate multi-label classifiers for predicting crystal structure in high entropy alloys. Cross-validation and alternative scoring functions are tools that can be used to optimize such models. Additional data on compositions with intermetallic phase identification could build these into more sophisticated classifiers. Although here we focused on experimental data, this approach could be applied to computationally-generated data in order to build such classifiers, which would then need to be validated with experimental results.

B. Property Prediction: FeCrAl Alloy Radiation Analysis

This case study focuses on predicting the yield strength of FeCrAl alloys subjected to neutron irradiation. The data come from tables published in a report documenting testing of FeCrAl alloys in the High Flux Isotope Reactor at Oak Ridge National Lab (ORNL) (Field et al. 2016). These alloys are of interest for accident tolerant fuel cladding due to their ability to form protective alumina scales while producing limited heat and H_2 during exposure to steam at elevated temperatures. Furthermore, alloying and oxide-dispersion strengthening are expected to allow tuning of mechanical properties for these harsh environments (Field et al. 2016).

The data reported by Field et al. include exposures of 11 different alloys to several different testing conditions, including pre-exposure (non-irradiated) conditions. The data were collected in two stages, with a first study that focused on composition effects “Phase I” and a second study that aimed to collect more information about radiation-induced changes in microstructure “Phase II.” As of the 2016 report, a total of 74 data points had been collected. Previously published work on the collected data found that composition effects were important in determining embrittlement outcomes, that hardening was correlated with phase-separation in high-Cr content alloys (Field et al. 2015; Field, Briggs, Sridharan, et al. 2017), and that dislocation loop formation occurred in a spatially heterogeneous manner that correlated with grain boundary position (Field, Briggs, Hu, et al. 2017). These results indicate that both composition and initial microstructure play a role in determining the embrittlement effects of FeCrAl alloys.

In the present work, we develop a simple regression model with regularized variable selection using the FeCrAl data that links composition and exposure variables to yield strength. We do not include microstructural variables because these were not available for individual observations in the dataset. The present discussion is meant only as a test study using the data available from the ORNL work, and is not meant as a critical analysis of that ongoing effort. Rather, we seek to identify ways in which a statistical-learning approach can enhance or extend the utility of such hard-won data, especially if additional researchers conducting similar tests elsewhere can contribute to overall data libraries. We also wish to articulate obstacles in the way of this extension and expansion and seek concepts for overcoming such obstacles.

The data were split into training and test groups, with the test data constituting 11 of the 74 data points. The remaining 63 data points were used for training with cross-validation. The 19 variables used for the analysis included estimated atomic fractions of 11 elemental species based on the weight-percent compositions reported in Field et al., irradiation temperature T_{irr}, test temperature T_{test}, irradiation dose d, four interaction terms (Cr/Al, Cr*T_{irr}, and Cr*d, and T_{irr}*d), and the estimated configurational entropy of the compounds based on the atomic fractions of species. Interaction terms were chosen because of the known importance of α' precipitation as an embrittlement driver in high-Cr containing alloys, a process that is triggered by exposure to irradiation and high temperatures. The Cr/Al ratio was included because both of these elements are typically associated with corrosion resistance, and an increase in one may couple to a decrease in the other, meaning that decoupling their effects could be difficult without an interaction term. Because the samples are not actually full solid solutions, the estimate of configurational entropy should be taken more as an indicator of the *potential* for solid solution, rather than as the actual entropy of the sample.

This is not an exhaustive list of variables that could have been included in the analysis. We could have included every compositional interaction (not just Cr/Al), every interaction between exposure and composition, and we could have also included powers of terms such as dose, as some amount of saturation as a function of dose has been noted in literature. Furthermore, we could have included other aggregating variables similar to the configurational entropy variable, such as melting temperature based on the rule of mixtures. Due to the limited number of data points, we decided to stick to the main effects and the few interaction terms that seemed pertinent.

In addition, we note that the use of atomic fractions as variables effectively limits a model to a particular alloy system with a particular set of impurities. Variables that can be derived from any composition, such as those used in the HEA case study, are a better choice when there is uncertainty about which atomic species may be present in future compositions of interest.

A prediction model for FeCrAl alloy yield stress was built using regularized linear regression on the 19 predictor variables. Data were centered and standardized prior to fitting in order to place all variables on the same scale. The model was built using the least angle regression algorithm with lasso regularization (lasso-LARS). The least angle method is an efficient algorithm for multivariate linear modeling that sequentially builds a model by including predictor variables one-by-one and scaling their coefficients to continually decrease correlations with residuals (Hastie, Tibshirani, and Friedman 2009). The lasso is a regularization method that penalizes models for using too many variables, and encourages scaling coefficients to zero. The size of the penalty is determined by α, a regularization parameter.

Figure 5 shows how the stepwise path of the lasso-LARS coefficients evolves as a function of the regularization parameter α, which controls the complexity of the fit. The optimal final value of α (0.0012) was chosen by cross-validation based on the cross-validated mean-square error of 9 validation folds. Of the initial set of 19 variables, 2 (Cr and N) are eliminated by the lasso algorithm, while 3 others (P, S, and S_{conf}) have coefficients of <100 (three orders of magnitude

below the largest coefficient). Initially, the Si content and test temperature stand out as the most strongly correlated variables, but by the end of the procedure the variables *d* (dose) and T_{irr}*d have the largest coefficients.

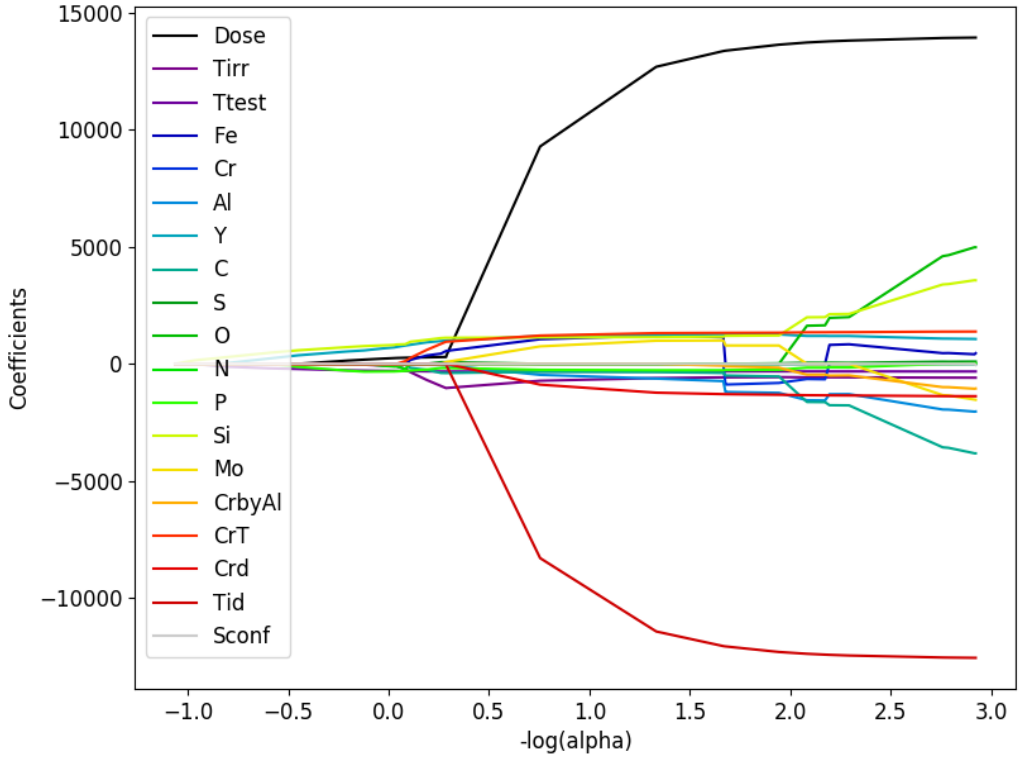


Figure 5. Coefficients fitting path of the lasso-LARS algorithm applied to FeCrAl alloy data for yield stress. Large positive coefficients indicate an embrittling effect, while large negative coefficients indicate a softening or toughening effect.

The final coefficients for the exposure variables reflect intuition: higher irradiation temperature or test temperature tends to have a softening influence, while large irradiation dosage leads to increased embrittlement. Similarly, increased Si, Y, O, and Fe seem to lead to embrittlement. The effect of oxygen is expected—oxygen is likely correlated with oxide precipitates in the material, which should have a strengthening and embrittling impact. Furthermore, increased impurity concentrations (e.g., Y, Si) should cause solid solution and further precipitate hardening. However, countering this intuition is the finding that the coefficients for Mo and C are negative, while the coefficient for Fe (i.e., definition of non-impurity) is positive. This could be an interaction effect (i.e., increased Fe will tend to imply decreased Al and Cr, and Al already has a negative coefficient), or it could indicate a microstructural effect, such as a shift in grain size, morphology, or precipitate formation and character when the primary alloying element concentration changes.

The Cr*T interaction is apparently an embrittling one; this is consistent with the observation that increased Cr content leads to α’ precipitate formation, and suggests that increased irradiation

temperature is necessary to facilitate precipitate formation. However, the Cr*d interaction shows a softening effect, which conflicts with the expectation that increased dosage accelerates α' precipitate formation. Meanwhile, the third interaction term in this triangle, T_{irr}*d, is very large, and it is negative (toughening). This likely balances against the coefficient for dose—because irradiation temperature and dosage are always 0 at the same time for these data, and increased dosage *alone* is an embrittling factor, the large negative coefficient for T_{irr}*d, along with the smaller negative coefficient for T_{irr} alone, balance the correlation of these variables. The linear model does not seem adequate to describe the true interactions between the processing variables and the Cr content of the alloys.

The final correlation between the fitted and true values for the training data, as shown in Figure 6, is fair, with an r^2 of 0.82. The average prediction error is 127 MPa (roughly 15%) when the model is applied to the test data. Analysis of the residuals indicates that for the most part the assumptions of the linear model are satisfied by the data; the residuals are fairly normally distributed (though heavy on the tails), and residual plots versus individual predictor variables do not show obvious deficiencies, aside from the fact that individual variables sometimes have obvious gaps in their domains, such as T_{irr}, d, and interactions of these. The plot of residuals versus fitted values (Figure 7) shows little dependence on the fitted yield stress.

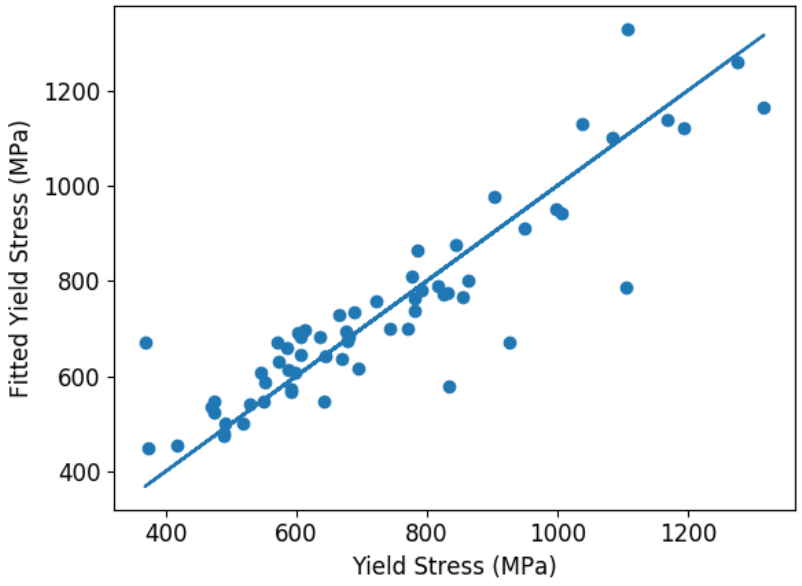


Figure 6. Fitted vs. actual values for FeCrAl yield stress training data

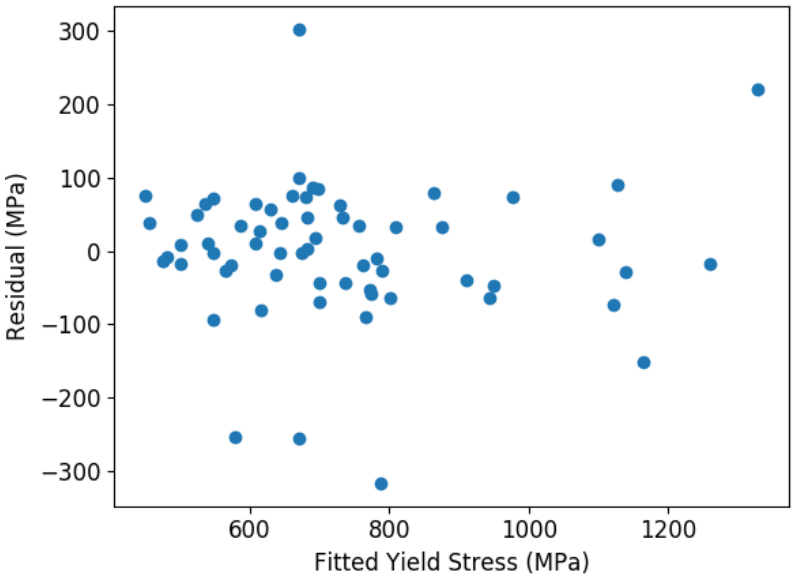


Figure 7. Residuals vs. fitted yield stress for the FeCrAl data

Missing information relevant to the embrittlement outcomes may exist, which could account for some of the variance that was not captured by the linear model. For example, the effect of grain boundary density (or grain size), which is known to influence dislocation formation, is absent from these data (Field, Briggs, Hu, et al. 2017). Similarly, the effect of α' precipitates is difficult to capture using composition and exposure variables alone; there are conflicting correlation effects from dose, temperature, and Cr content, and details like grain size, impurity content, and dislocation density could influence the formation kinetics of precipitates. This is a case where specific modeling of precipitate formation coupled to either simulations or automated image analysis could tell us which compositional interaction terms or microstructural variables might be most important for predicting embrittlement of a particular alloy upon exposure to high temperatures and neutron irradiation.

Indeed, this case study finds that a more in depth analysis with additional data is needed in order to gain mechanistic insight into radiation-induced FeCrAl alloy embrittlement. We should not expect compositional variables alone to predict yield stress, even at high temperatures, because yield is a process that inherently involves material microstructure. Furthermore, while composition should not be expected to remain constant throughout exposure to high temperatures due to processes like oxidation and corrosion, the effect of irradiation that leads to embrittlement is primarily microstructural alteration. The process of irradiation-induced microstructural change is heavily governed by kinetics, which suggests that compositional and environmental variables alone cannot capture the full range of important predictors, especially without incorporating interaction terms. Such physical intuition is required to build a more effective model.

Field et al. engaged in extensive microstructural analysis of subsets of their samples, including dislocation loop analysis and α' precipitate analysis (Field, Briggs, Hu, et al. 2017; Field

et al. 2016). Volume fraction, distribution, and number density of microstructural features are important intermediate variables that can be used to build PSP linkages, but initial microstructural variables such as grain size, porosity, and precipitate distribution are also needed. Work to this effect at ORNL has employed automated image analysis algorithms in the past to characterize microstructural features in FeCrAl alloys (Field, Briggs, Hu, et al. 2017; Field et al. 2016). However, these efforts have met with limited success, still requiring manual counting techniques to distinguish among various dislocation orientations (Field, Briggs, Hu, et al. 2017).

There are limits to our ability to build microstructural databases, even with automated image quantification. Specimen preparation, especially for TEM or 3D characterization, can be extremely time consuming and require highly specialized training. Image analysis algorithms have limitations—supervised algorithms can only detect what they are trained to detect, and they must avoid being tricked by image contrast artifacts that may be easily identified by experts. Nonetheless, in cases where sample preparation and image collection are well-tuned, and where features of interest are either distinctive or difficult for a human analyst to quantify, automated analysis can be an effective, low-bias tool that maximizes the value of each image. It can also be useful in the context of non-destructive evaluation, where inspections aided by computer image recognition are already of interest (Brence and Brown 2002).

Finally, we note that this is a case where a computationally-generated microstructure could have real value, especially when coupled to models of microstructure evolution or simulation of mechanical properties. Tools like DREAM.3D and PyMKS may be useful for such analysis, where the effect of precipitate size, shape, or distribution could be modeled.

C. Microstructural Analysis

This case study is aimed at exploring some aspects of automated image analysis using micrographs taken from the ASM Micrograph Database (Lupulescu et al. 2015). Specifically, 11 optical images of flake and lamellar grey cast iron were analyzed using different methods of edge detection and segmentation to understand some of the strengths and weaknesses of these approaches. These images were chosen because of their distinctive microstructure that features graphitic flakes or lamellae that appear as dark regions against a much lighter background of primarily pearlite matrix. Depending on the composition of the cast iron, the size, shape, and density of the graphitic phase varies. In fact, the graphitic phase can be either hypoeutectic or eutectic and the matrix can contain both ferrite and pearlite Fe, but our image analysis does little to distinguish between these phases. Four of the 11 images were listed as “unetched” in the database, while the others were etched with 4% picral. The difference in etch significantly changes the contrast of the images for the various phases, though graphitic regions still appear dark and the matrix still appears light. Otherwise, the samples varied in terms of composition, specifically with regard to C, P, or Si content.

The goal for this case study was to use computational image analysis methods to quantify information about the shape, size, and quantity of the graphitic phase in the set of images. The first

task in this analysis was to segment the images so that the graphitic regions would be identified separately from the matrix phase. A few methods were attempted including canny edge detection followed by edge filling and masking followed by Sobel or Laplace edge detection.

A side-by-side comparison of these segmentation procedures is revealing. Canny edge detection fails to effectively fill single flakes due to discontinuities in detected edges and edge detection at small contrast variations internal to flakes. In contrast, the watershed algorithm is much more successful. In this process, thresholds are first applied to an image to define regions that correspond to particular objects with high confidence (i.e. the lightest regions are almost certainly matrix and the darkest regions are almost certainly graphitic). Then, contrast gradients are identified in the image using a gradient detection method such as the Sobel (1^{st} derivative) or Laplace (2^{nd} derivative) transforms. Following this, objects are filled in by extrapolating contrast along gradients, in the same way that one might try to predict water running down hills on a topographic map (i.e., the watershed algorithm). The Laplace and Sobel transforms have differing strengths. Laplace edge detection is more successful in delineating smaller flakes, but also exhibits higher sensitivity to the choice of initial thresholds as shown in Figure 10.

Figure 10 compares the estimated phase fraction of the graphitic flakes and lamellae of one example image with composition 4.0% C and 2.0% Si using the watershed algorithm with the Laplace and Sobel transforms. A range of high and low threshold values was applied in order to determine how sensitive the two techniques were to the choice of thresholds. This can also be seen as an assessment of the sensitivity of this algorithm to brightness/contrast variations from different microscopes or imaging technicians. Figure 10 shows that the Sobel transform exhibits more systematic variation in the estimated phase fractions as a function of threshold choice as compared to the Laplace transform. The Sobel transform also shows much smaller variance overall. The Laplace transform consistently estimates a higher graphitic phase fraction, presumably due to its ability to more adequately detect small features. It turns out that the Laplace transform is also more easily confounded by contrast variations in the matrix of the picral-etched images. Further analysis was therefore conducted using the watershed algorithm with the Sobel transform.

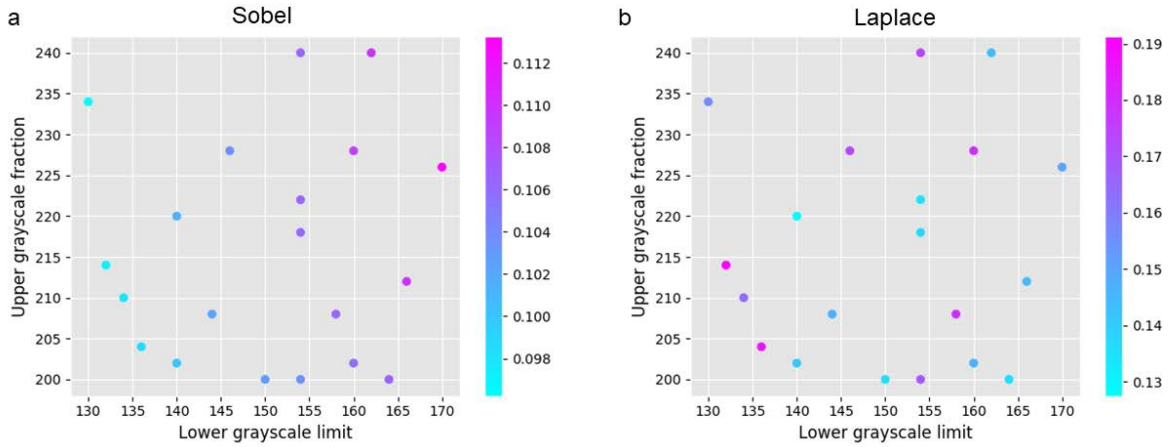


Figure 8. Estimated graphitic phase fraction (color bar) as a function of thresholding limits for the watershed algorithm using the Sobel (a) or Laplace (b) transforms

After segmenting an image, the next step is to analyze the objects in the image. If objects do not overlap, it is fairly straightforward to simply define objects as those occupying a continuous space with the same segmentation label (graphite or matrix). If they do overlap, methods must be employed to separate overlapping objects for analysis. These were not explored in the present study.

Once image objects have been identified, their properties can be easily quantified. In the present work, this was done using the scikit-image function measure.regionprops, which can determine a large number of object properties including area, orientation, eccentricity, solidity, and perimeter (van der Walt et al. 2014). It is then possible to construct histograms of the object properties in an image to understand how the shape, size, and other features of objects are distributed. Objects can also be filtered using their properties; for example, segmentation errors may appear as abnormally large area flakes. In images of other materials, one could imagine distinguishing between pores and grain boundaries using shape descriptors.

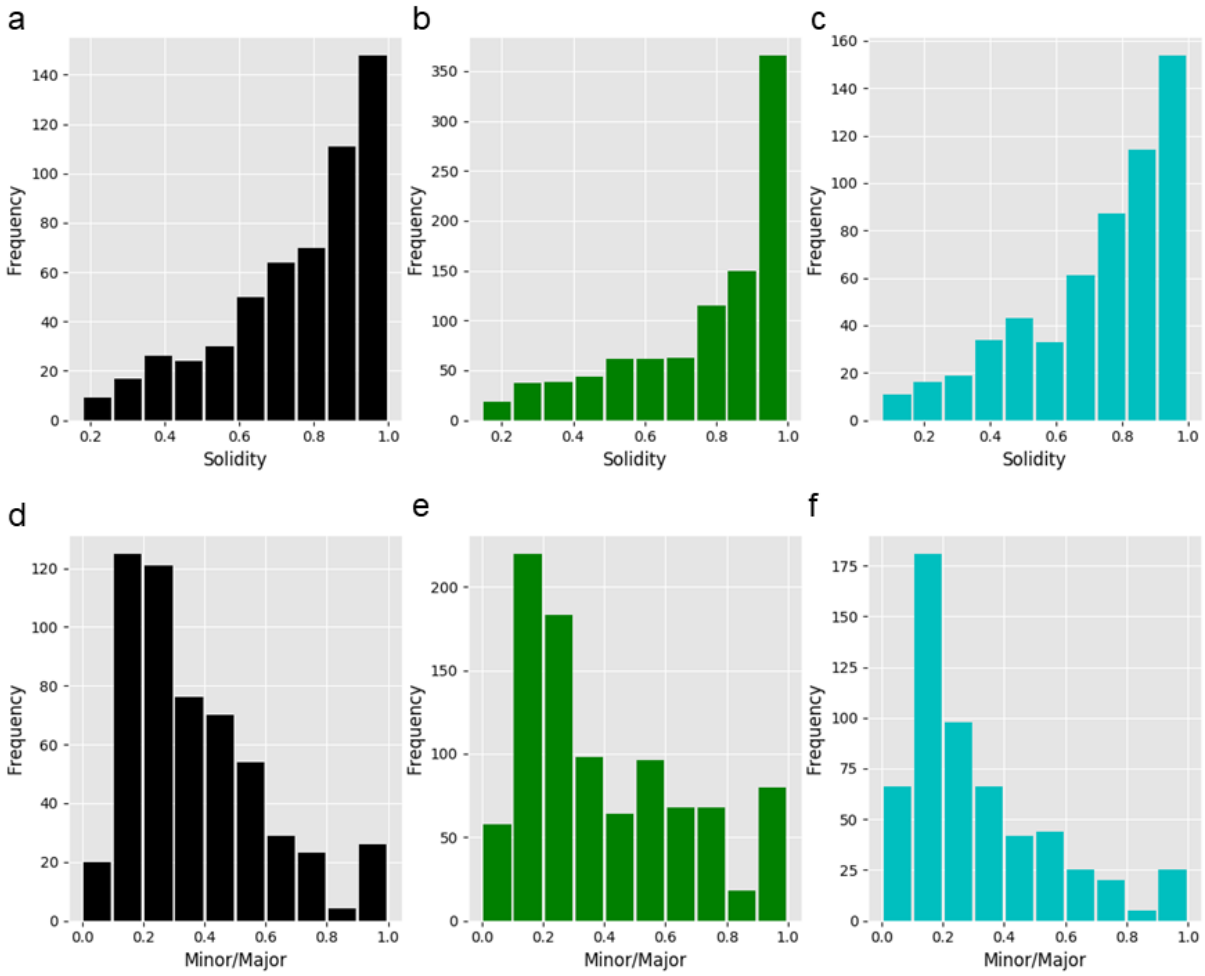


Figure 9. Example histograms for graphitic flake solidity and minor/major axis ratio for three images of grey cast iron. From left to right, the compositions are: (a,d) 2.1% C, 2.0% Si (b,e) 3.0% C, 2.1% Si, (c,f) 3.7% C, 2.1% Si.

Figure 11 shows example object property histograms for three images of unetched grey cast iron with differing composition. Specifically, these histograms show solidity (ratio of object area/convex hull area) and the ratio of the minor and major axis, defined for each object by the ellipse with the same normalized second moment. These histograms show, for example, that the sample with 3.0% C has a much higher number density of graphitic flakes compared to the other two images. On average, the graphitic flakes in the sample with the highest C content have the least solidity (ratio of object area/convex hull area) and are the most slender in character. Although not shown, there is also an increase in the mean and standard deviation of flake area with increased C content for the three images.

A second way to characterize microstructures in images like these is to use n-point statistics. Rather than focusing on the shape and size properties of individual objects, this approach is more useful for understanding distributions in space of objects in an image. Such analysis can convey information about object shapes, but it will do so in an averaged way.

After segmenting the images using the watershed algorithm with the Sobel transform (threshold pixel intensities [8-bit] at 90 and 190 for picral-etched images, and 150 and 220 for unetched images), the images were processed using PyMKS to identify their n-point statistics. The 1-point statistic of the images defines the estimated graphitic phase fraction. Figure 12 shows this estimated graphitic flake phase fraction f_g as a function of C content for the 11 images. Note that all four points with apparent f_g below 10% are from unetched images (points with black outline); the same compositions after picral etch and corresponding image processing show the same trends in f_g as a function of C content, but shifted to higher absolute estimated f_g.

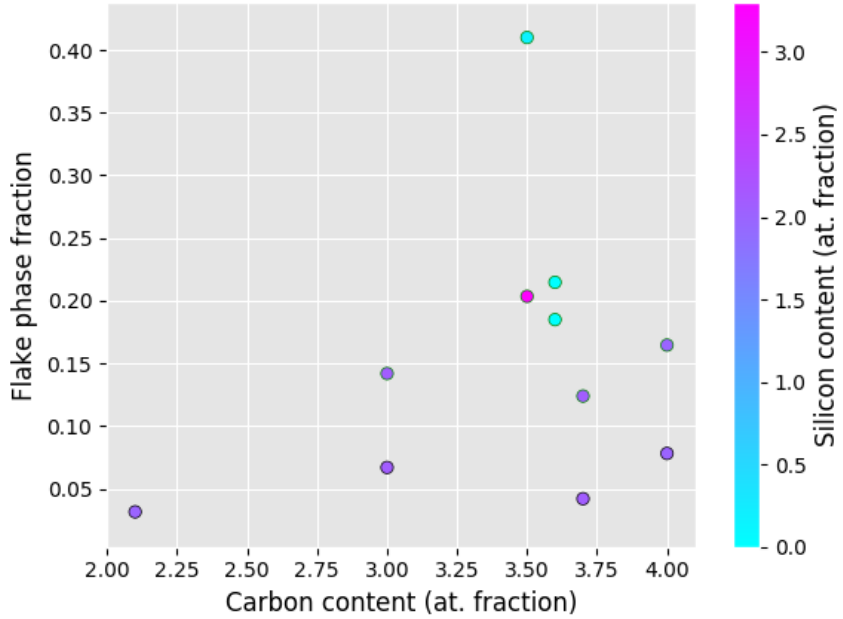


Figure 10. Estimated flake phase fraction based on n-point statistic image processing as a function of C content. Points outlined in black are from unetched images. Data point color corresponds to Si content.

Principal component analysis applied to the n-point statistic outputs of the images showed that the etch type was a major source of variance. Figure 13 shows that the four unetched images were clearly separated from the other seven images in both PC1 and PC2. The role of composition in principal component space is less clear.

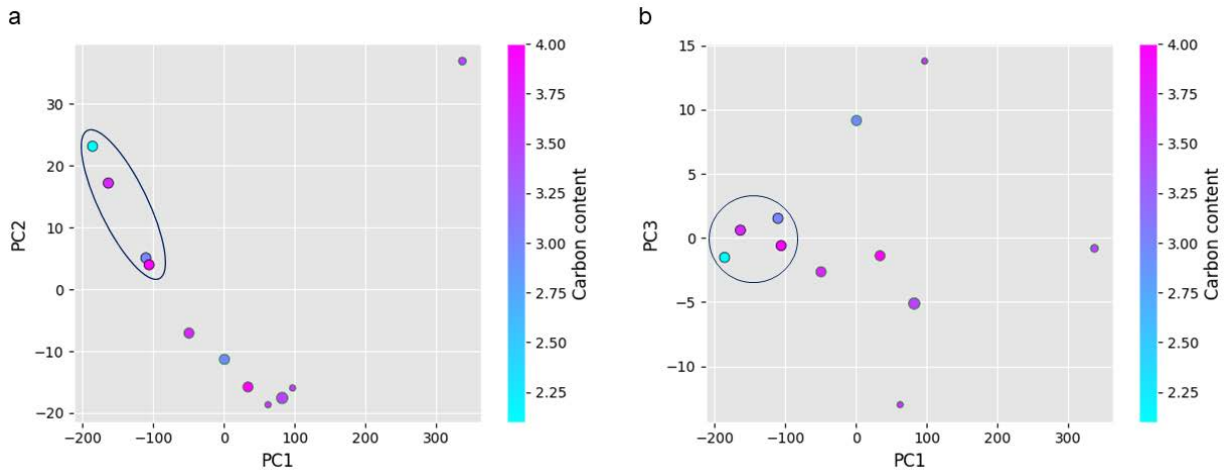


Figure 11. Results of principal component analysis on grey cast iron images converted to n-point statistic representations. Circled points correspond to images of unetched samples.

The above analysis exemplifies some of the pitfalls of automated image processing. Before developing an automated workflow, it is essential to understand the expected error due to effects like varied brightness/contrast, choice of etchant, choice of edge detection or segmentation method, and unexpected phase variations. Even among the images analyzed here, additional phases were present that were not detected using these methods, including precipitates that appeared in samples containing P. While databases of image analysis results may be an effective substitute for databases of actual images, the analysis supporting such results needs to be validated and documented so that potential users and contributors understand these limitations.

Human image analysis has its own advantages and limitations. For example, when several PhD materials scientists were asked to describe one of the images of grey cast iron, all noted that many of the graphite flakes emanated from central points in a star-like pattern. This observation stands out immediately to the human eye, but would be difficult to find with an automated code that was not specifically tailored to identifying such features. However, human analysts were not able to easily provide quantitative descriptions of flake size or morphology that can be acquired through computational code.

An image of reaction-bonded Si-C was also shown to two of these materials scientists, who were challenged to provide qualitative descriptions of the images and to attempt quantitative analysis by whatever methods they preferred. They were told that this was an optical micrograph of reaction-bonded Si-C, and were given the etch information provided by the ASM database. While both identified three distinct phases in the image, there was disagreement about the identity of the dark gray phase, including some uncertainty about the extent to which imaging or microstructural defects were responsible for the darker shade. One participant used manual point-counting and circle-intersection methods to quantify the microstructure (Boyer and Gall 1981), while the other used computational image analysis tools (specifically, the Trainable Weka Segmentation Plugin with the FIJI distribution of ImageJ [Arganda-Carreras et al. 2017]).

Estimates by these two methods of the Si phase fraction (white phase) were 12–13% and 18%, respectively. Estimates of the SiC phase fraction (light phase) were 55–58% and 43%, respectively. The 43% estimate was noted by the analyst as likely an underestimate, due to the difficulty of separating very small particles of the light gray phase from dark gray surroundings. The first participant also estimated grain sizes manually (mean ~12 um for the light gray phase), while the second participant generated a particle-size distribution from the image segmentation analysis. Neither participant identified any information about feature shapes (e.g., aspect ratio) or spatial distributions, though this may have been due to limited time.

This discussion demonstrates that image analysis conducted by individuals can give different outcomes, and that different features may be interpreted differently, noticed, or ignored, by different people. Furthermore, lack of metadata can strongly affect the interpretation of images. For example, the grey cast iron image series in fact includes up to four phases (pearlite, ferrite, hypoeutectic graphite, eutectic graphite), but these are not always obvious to an observer without additional characterization (e.g., X-ray diffraction results), processing information, or higher-magnification images. Image analysis requires understanding of both *methods* and *subject*—computational analysis has real potential to accelerate manual methods (which are very often systematic and can be straightforward to automate), but only if conducted with appropriate context and awareness of potential pitfalls. The ability to process a larger volume of images per sample also has the potential to produce improved statistics to represent a 3D microstructure based on 2D slices.

Quantitative image processing and analysis is not a staple of current materials science curricula at universities. A review of the materials science course catalogs of 12 universities[5] found only two containing courses specifically targeting the overlap of statistical and informatics techniques with materials science.[6] At Carnegie Mellon, in particular, data science was highlighted as a topic covered in at least three courses. Most of the universities do provide courses on imaging methods (usually more focused on understanding the function and appropriate use of various types of microscopes, especially transmission electron microscopes, than on understanding image post-processing and analysis or microstructure quantification) and computational materials science (typically focused on computational material modeling approaches, such as density functional theory, kinetic Monte Carlo simulations, or molecular dynamics simulations). These types of courses can provide a natural foothold for adding topics on image analysis techniques and materials informatics approaches to modeling materials data. Materials science departments can also recommend courses taught in other departments on statistical learning, data mining, and image

[5] MIT, UC Berkeley, University of Illinois Urbana-Champaign, University of Michigan Ann Arbor, Georgia Tech, Stanford University, Northwestern University, Cornell University, UC Santa Barbara, Penn State University Park, Caltech, and Carnegie Mellon University

[6] Cornell University: MSE 5730 “Probability, Statistics, and Data Analysis for the Physical Sciences” and Carnegie Mellon University: 27-566 “Special Topics in MSE: Using Materials Informatics to Assess Societal Impact of Materials”

analysis. For example, the biology and bioengineering communities use these tools extensively, and there is substantial overlap between the tools and techniques to solve image analysis problems in biological sciences and materials science (Russ 2011). Courses on image processing and analysis already exist in some biological and bioengineering departments,[7] which may be useful models for designing similar courses focused more on materials science. The first step in promoting these tools for future materials research and development will be to make them visible to students in universities.

[7] Stanford, Biomedin 260: "Computational Methods for Image Analysis and Interpretation", University of Chicago, 2019 offering: "Computational Image Analysis in Cellular and Developmental Biology", MIT 6.121: "Biomedical Signal and Image Processing"

6. Conclusions and Recommendations

This report covered a broad spectrum of topics in order to build a picture of the current status, limitations, and opportunities of data science and statistical learning methods for understanding, modeling, and designing materials for extreme environments. We conclude with key observations on areas where these approaches could be of the most immediate value, with some comments on barriers to adoption and the necessary shifts in institutional policy and workforce training needed to facilitate this kind of work.

A. High Value Challenges for Materials in Extreme Environments

A single, universal database that applies to all materials of interest and incorporates all relevant experimental and computational results in a validated fashion is still a long way away. However, there are some specific challenges that we can begin to solve using the tools of data science.

First among these is the search for high entropy alloys and ceramics with properties optimized for high temperature, corrosion resistance, radiation resistance, and tolerance of high strain rates. Because of the early stage of CCA/HEA and high-entropy ceramic research, there is real opportunity to develop databases that relate measured properties, composition, processing, and microstructure. We saw in the case studies that even limited data on HEAs can be used to build decent classification schemes—a more complete database accounting for microstructure in addition to composition could accelerate work targeting specific applications, properties, or mechanistic insights. Work on building computationally-generated databases on HEAs is already under way, but augmenting such data with experimental results is essential and will benefit the entire HEA/CCA community. Even within the DoD, establishing a common reporting architecture to be shared among DoD labs would be worthwhile.

More ambitious would be developing a resource similar to the TE Design Lab that allows reporting and curation of both experimental and computational data and provides tools for analysis and visualization. Similar efforts to build shared resources in specific areas of materials for extreme environments (corrosion, radiation, high temperature, hypersonic) seem more likely to be successful than simply uploading results into universal data repositories. The DOE has funded development of such tools in the past—it may be time for the DoD to do the same. Short of funding tool development or data management/curation, the DoD would benefit from providing an internal materials resource registry searchable by focus area. Resources built and hosted by individual groups in the various extreme environment research communities could then be linked to this

registry, in addition to potentially being linked to larger repositories such as the NIST Materials Resource Registry.

With data available, analysis techniques for building exploratory, explanatory, and possibly utilitarian models could be developed with some of the techniques highlighted in this report. Frameworks that accommodate multi-label classification or fuzzy logic will likely be of great value in understanding materials data. Methods with explanatory power, such as decision trees and regression, will be valuable for building materials knowledge and wisdom. However, more opaque techniques like neural nets should not be discounted, especially when there is a highly utilitarian goal with multiple, complex, or poorly established mechanisms, such as predicting corrosion in disparate materials as a function of exposure conditions and material composition, microstructure, and loading.

There are several areas in which automated, computational image analysis can make a real impact for the design, monitoring, and development of materials in extreme environments. Because automated image analysis can quantitatively and consistently describe material microstructure within a particular material system (e.g., through 2-point statistics or feature distribution functions) it can facilitate robust comparison among multiple images without relying on consistent human interpretation. Therefore, images taken by different operators, at different facilities, or at different times, but with the same general imaging subject, conditions, and methods, could be compared more readily and quantified more consistently. Furthermore, such methods significantly enhance the speed of image characterization, meaning that a statistically representative description of material microstructure can be gathered easily once an imaging and analysis method has been established. Finally, computational image characterization can condense image files into more compact representations of relevant features, decreasing the electronic storage space needed to maintain a database including such information.

Perhaps the most useful place for such analysis in the extreme environment world is for quantitative microstructural characterization alongside ongoing exposure to extreme environments. This example applies to both research facilities and industrial settings. The reactor embrittlement databases are ideal examples of the kind of work that can be augmented by such characterization (IAEA 2009; Wang and Subramani 2008). Reactors are installed with test coupons ready for periodic characterization for monitoring pressure vessel health. Coupons are regularly tested for mechanical properties and material property changes are tracked over time. If those same coupons were also characterized with a few standardized scanning electron and optical microscope conditions, the resulting data could inform modeling and simulation efforts to predict radiation damage (and ultimate failure) or understand the mechanistic causes of embrittlement. It is easy to see how such an approach could be extended to studying corrosion, oxidation, or high temperature microstructure evolution. Where it is impractical to include test coupons in operating systems, such concepts could be extended to non-destructive evaluation (NDE) methods such as eddy current testing. Automated analysis results could be used in conjunction with human analysts as they make decisions about the severity of material flaws (Brence and Brown 2002). This kind of

ongoing material microstructural monitoring has the potential to enhance the sustainment and readiness missions of the DoD.

A second important area of impact for these methods lies in material processing and quality assurance. In a manufacturing setting, lots of identically produced material or parts are generally tested to ensure that they meet quality assurance standards. A sample will be extracted for testing, and measured properties must fall within a specified range that is defined for each product. If the sample falls outside of that range, it can be an indication of a problem in the processing line. At minimum, the lot will be rejected or rated at a lower grade. At maximum, the processing line will be inspected to identify the source of the problem and correct it. In particular, fiber composites for hypersonic systems stand out as an area where such computational image analysis could augment manufacturing quality assurance (QA) efforts, and potentially assist in improving overall manufacturing consistency.

QA is an ideal case for automated image analysis, as the samples and test conditions should be identical from lot to lot, even across multiple manufacturing facilities or QA technicians. Therefore, it should be straightforward to define tolerance ranges based on automated characterization workflows that quantify the key microstructural features of interest. Such concepts are already being applied in the world of additive manufacturing, where researchers are training neural networks on videos from the build chamber in order to predict build quality (Metal AM 2018). QA is an essential aspect of any manufacturing plant, both to maintain standards for customers and to optimize process control that in turn manages maintenance costs and surprises. An increase in material grain size or a change in precipitate density or porosity can be informative of potential drift in a manufacturing line, for example, pointing to temperature control failures or changes in impurity levels. Furthermore, if a lot passed QA testing but later returned with failures before the expected end of service life, information about the starting microstructure could be valuable in understanding the cause of the failure. Since many extreme-environment damage processes are slow moving, hard to monitor, and poorly modelled, there are limits to what QA testing can do to predict extreme-environment related failures. However, this does not mean that such data would not be useful in aiding continuous improvement of processing lines.

Characterizing expected variance in analysis results due to normal deviations in imaging conditions and material microstructure is an important task in any attempt to use computational image analysis to quantify material microstructure. Furthermore, it will be necessary to precisely define the analysis workflow that will output the variables of interest. Standardized software tools that can run such workflows in an automated fashion, include documentation of validation and tool limitations, and provide guidance for analysts on how to manage various types of imaging artifacts will substantially improve the reliability of such efforts across multiple analysts or facilities.

B. Implementation Challenges and Workforce Development

There are several things that need to happen before statistical learning approaches can be widely applied to problems in materials for extreme environments relevant to the DoD. These

relate to the availability and discoverability of data and analysis tools, the standardization of data storage, sharing, and curation, and the availability of a workforce that is skilled in the application of these methods. The responsibility for providing these pieces is shared between DoD leadership, research communities, professional societies, and academia.

The first hurdle is to establish a culture of database-building with standardized formats for data storage and data entry. Individual researchers can publish their data entry formats, which can potentially smooth the process of sharing data when labs choose to collaborate. In cases where a small database has been collated for years, linking that database to a resource registry can also promote similar practices by those collecting similar data. While straightforward in principle, however, it is difficult for DoD researchers to do even this much without institutional support. It takes time, effort, and expertise to load data onto institutional servers in formats that are searchable and well-referenced. Furthermore, information access must be managed through approval processes and access to useful software may be restricted or outdated; these are both significant barriers to building such resources.

As described in chapter 3D, institutional support can come in a number of forms, from providing guidance on the best practices for data storage and sharing, to supporting committees that establish data storage standards, to directly contracting for development of e-collaboration tools. A more in depth analysis of the data management practices of the materials research and testing communities of the DoD could provide more specific steps forward.

We do not recommend that the DoD attempt to establish its own broad materials data repository, as this would simply duplicate efforts already underway by NIST and the Materials Genome Initiative. Rather, we suggest that smaller, topic-centric databases would be of greater value to the DoD materials community. The area of CCAs/HEAs is the best one for testing out a materials data storage paradigm that incorporates not only composition and property information, but also the results of microstructural analysis. Such a database would require substantial collaboration between materials researchers and experts in data management, software development, and image processing. The result could lead the way and be extended to other communities within the DoD, such as those focusing on hypersonic materials, corrosion, and radiation.

In addition to the process of database-building, sharing of analysis tools and protocols across DoD labs could be of value. For example, if one lab develops an open-source tool that detects microstructural features (e.g., pore sizes and shapes), that lab can make that tool available to other DoD researchers, even if specific data cannot be shared. With such efforts, it will be important to validate tools in order to understand likely sources of error, expected variance, and applicable ranges.

Finally, it is necessary to build awareness and technical skill with data science and statistical learning approaches within the materials science workforce. Universities need to inform students of available databases, data repositories, and statistical learning tools, and provide some

background as to how they might be useful to the materials community. Furthermore, computational image analysis and quantification should be incorporated into materials science course catalogs. Practical coverage of available image characterization software, useful image processing methods (e.g., edge detection and segmentation), and statistical approaches to interpreting and describing imaging data (e.g. n-point statistics) would be valuable to any materials experimentalist's toolbox. Courses covering these topics could be designed based on courses now available to the biology and bioengineering communities. Similar courses in materials curricula would enhance workforce literacy in these areas and speed the implementation of such tools in relevant manufacturing, research, or monitoring facilities. University libraries and professional societies could further assist this process by publicizing the existence of various materials data science tools, including databases, repositories, and software.

DoD researchers in the area of materials for extreme environments can benefit from applying data mining and statistical learning tools, including those that permit performance classification, property prediction, and microstructural quantification. These have the potential to help the materials research and testing community build models that can estimate appropriate environments for material use, predict material failure, optimize material structure, and make data-driven decisions about where to focus experimental and measurement efforts. However, this will only be possible with access to appropriate tools, institutional support, and a skilled workforce.

Appendix A.
Selected Materials Data Science Links

- Materials Genome Initiative: https://www.mgi.gov and https://mgi.nist.gov
    - PRISMS – Predictive Integrated Structural Materials Science: https://www.prisms-center.org
    - Materials Commons: https://materialscommons.org/mcapp/#/data/home/top
    - Materials Project: https://materialsproject.org
    - CHIMAD – Center for Hierarchical Materials and Design: chimad.northwestern.edu
- NIST
    - Materials Resource Registry: https://materials.registry.nist.gov
    - Materials Data Repository: https://materialsdata.nist.gov
    - Materials Data Facility (funded by NIST and CHIMAD): https://materialsdatafacility.org
    - Standard Reference Data: https://www.nist.gov/srd/srd-catalog
- MATIN – Materials Innovation Network: https://matin.gatech.edu
    - PyMKS –The Materials Knowledge System in Python: pymks.org
- OQMD – The Open Quantum Materials Database: oqmd.org
- AFLOW – Automatic-FLOW for Materials Discovery: aflowlib.org
- MatNavi – NIMS Materials Database: mits.nims.go.jp/index_en.html
- ICSD – International Crystal Structure Database – www2.fiz-karlsruhe.de/icsd_home.html
- ASM International Online Databases: https://www.asminternational.org/materials-resources/online-databases
    - ASM Alloy Center Database
    - ASM Alloy Phase Diagram Database
    - ASM Failure Analysis Database
    - ASM Handbooks Online

    - ASM Medical Materials Database
    - ASM Micrograph Database
    - Corrosion Analysis Network
- TE Design Lab – Thermoelectrics Design Lab: www.tedesignlab.org
- Registry of Research Data Repositories: re3data.org
- Mat-DB: https://odin.jrc.eu.europa.eu/alcor/Main.jsp
- Scikit-image –image processing in Python: https://scikit-image.org
- OpenCV: https://opencv-python-tutroals.readthedocs.io/
- DREAM.3D: dream3d.bluequartz.net
- Brookhaven National Laboratory – Center for Data-Driven Discovery: https://www.bnl.gov/compsci/C3D/

Appendix B.
List of Contacts

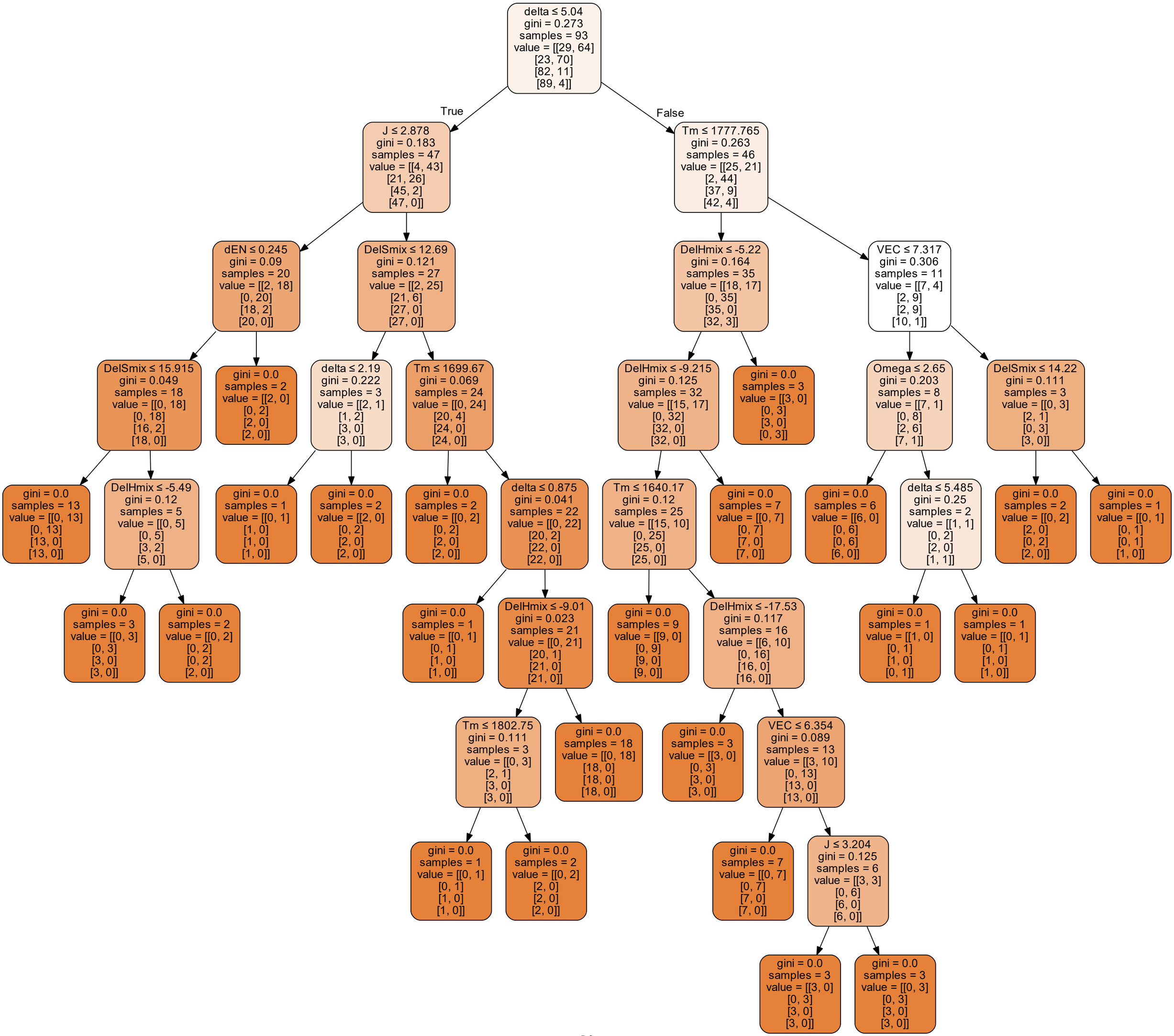
- Joel McDuffee, Distinguished R&D Staff and Group Leader, Nuclear Experiments and Irradiation Testing Group, Oak Ridge National Laboratory
- Dr. Farrel J. Martin, Materials Engineer, Code 6134, Center for Corrosion Science and Engineering, U.S. Naval Research Laboratory, Washington DC, 20375
- Jay Ong, Naval Surface Warfare Center, Carderock
- Dr. James Smialek, Senior Technologist, Materials High Temperature Oxidation, NASA Glenn Research Center
- Dr. Eric Wuchina, Program Officer, Ultra-High Temperature Materials, Office of Naval Research
- Dr. William Hong, Dr. Evan Laprade, Dr. Howard Last, Dr. Christopher Pellegrinelli, Dr. Janet Sater, Institute for Defense Analyses

Appendix C.
HEA Decision Tree

This appendix includes details for interpreting the high entropy alloy (HEA) decision tree built as described in chapter 5A. Table 6 explains the variable symbols in the tree. Each node in the tree lists the split condition (variable and value), the Gini impurity of the node before the split, the total number of data points (samples) to be split, and the label matrix (value) of those data points. The rows of the label matrix correspond to the class labels in the following order: FCC, BCC, σ phase, α phase. The columns represent binary class membership for that label in the following order (0, 1), with 1 representing a positive label. So, looking at the first node in the tree, there are a total of 93 samples, with an initial Gini index of 0.273. Of the 93 samples, 64 contain the FCC phase, 70 contain the BCC phase, 11 contain the σ phase, and 4 contain the α phase. The split condition for the node is average atomic deviation less than or equal to 5.04 %; if this condition is met, we move to the node on the left, and if it is not, we move to the node on the right. The same procedure is followed for all following nodes, using the new node split condition and the reduced set of data points remaining at that node.

Table 6. Symbols Used in the Final Decision Tree

Symbol (Tree)	Symbol (text)	Units	Description
delta	δ	%	Average atomic deviation
J	J	unitless	Average total angular momentum quantum number
Tm	T_m	K	Melting point
DelSmix	ΔS_{mix}	J/(mol*K)	Entropy of mixing
DelHmix	ΔH_{mix}	kJ/mol	Enthalpy of mixing
dEN	$\Delta\chi$	unitless	Electronegativity deviation parameter
VEC	VEC	unitless	Valence electron concentration


delta ≤ 5.04
gini = 0.273
samples = 93
value = [[29, 64]
[23, 70]
[82, 11]
[89, 4]]
True
False
J ≤ 2.878
gini = 0.183
samples = 47
value = [[4, 43]
[21, 26]
[45, 2]
[47, 0]]
Tm ≤ 1777.765
gini = 0.263
samples = 46
value = [[25, 21]
[2, 44]
[37, 9]
[42, 4]]
dEN ≤ 0.245
gini = 0.09
samples = 20
value = [[2, 18]
[0, 20]
[18, 2]
[20, 0]]
DelSmix ≤ 12.69
gini = 0.121
samples = 27
value = [[2, 25]
[21, 6]
[27, 0]
[27, 0]]
DelHmix ≤ -5.22
gini = 0.164
samples = 35
value = [[18, 17]
[0, 35]
[35, 0]
[32, 3]]
VEC ≤ 7.317
gini = 0.306
samples = 11
value = [[7, 4]
[2, 9]
[2, 9]
[10, 1]]
DelSmix ≤ 15.915
gini = 0.049
samples = 18
value = [[0, 18]
[0, 18]
[16, 2]
[18, 0]]
gini = 0.0
samples = 2
value = [[2, 0]
[0, 2]
[2, 0]
[2, 0]]
delta ≤ 2.19
gini = 0.222
samples = 3
value = [[2, 1]
[1, 2]
[3, 0]
[3, 0]]
Tm ≤ 1699.67
gini = 0.069
samples = 24
value = [[0, 24]
[20, 4]
[24, 0]
[24, 0]]
DelHmix ≤ -9.215
gini = 0.125
samples = 32
value = [[15, 17]
[0, 32]
[32, 0]
[32, 0]]
gini = 0.0
samples = 3
value = [[3, 0]
[0, 3]
[3, 0]
[0, 3]]
Omega ≤ 2.65
gini = 0.203
samples = 8
value = [[7, 1]
[0, 8]
[2, 6]
[7, 1]]
DelSmix ≤ 14.22
gini = 0.111
samples = 3
value = [[0, 3]
[2, 1]
[0, 3]
[3, 0]]
gini = 0.0
samples = 13
value = [[0, 13]
[0, 13]
[13, 0]
[13, 0]]
DelHmix ≤ -5.49
gini = 0.12
samples = 5
value = [[0, 5]
[0, 5]
[3, 2]
[5, 0]]
gini = 0.0
samples = 1
value = [[0, 1]
[1, 0]
[1, 0]
[1, 0]]
gini = 0.0
samples = 2
value = [[2, 0]
[0, 2]
[2, 0]
[2, 0]]
gini = 0.0
samples = 2
value = [[0, 2]
[0, 2]
[2, 0]
[2, 0]]
delta ≤ 0.875
gini = 0.041
samples = 22
value = [[0, 22]
[20, 2]
[22, 0]
[22, 0]]
Tm ≤ 1640.17
gini = 0.12
samples = 25
value = [[15, 10]
[0, 25]
[25, 0]
[25, 0]]
gini = 0.0
samples = 7
value = [[0, 7]
[0, 7]
[7, 0]
[7, 0]]
gini = 0.0
samples = 6
value = [[6, 0]
[0, 6]
[0, 6]
[6, 0]]
delta ≤ 5.485
gini = 0.25
samples = 2
value = [[1, 1]
[0, 2]
[2, 0]
[1, 1]]
gini = 0.0
samples = 2
value = [[0, 2]
[0, 2]
[2, 0]
[2, 0]]
gini = 0.0
samples = 1
value = [[0, 1]
[0, 1]
[0, 1]
[1, 0]]
gini = 0.0
samples = 3
value = [[0, 3]
[0, 3]
[3, 0]
[3, 0]]
gini = 0.0
samples = 2
value = [[0, 2]
[0, 2]
[2, 0]
[2, 0]]
gini = 0.0
samples = 1
value = [[0, 1]
[0, 1]
[1, 0]
[1, 0]]
DelHmix ≤ -9.01
gini = 0.023
samples = 21
value = [[0, 21]
[20, 1]
[21, 0]
[21, 0]]
gini = 0.0
samples = 9
value = [[9, 0]
[0, 9]
[9, 0]
[9, 0]]
DelHmix ≤ -17.53
gini = 0.117
samples = 16
value = [[6, 10]
[0, 16]
[16, 0]
[16, 0]]
gini = 0.0
samples = 1
value = [[1, 0]
[0, 1]
[1, 0]
[0, 1]]
gini = 0.0
samples = 1
value = [[0, 1]
[0, 1]
[1, 0]
[1, 0]]
Tm ≤ 1802.75
gini = 0.111
samples = 3
value = [[0, 3]
[2, 1]
[3, 0]
[3, 0]]
gini = 0.0
samples = 18
value = [[0, 18]
[18, 0]
[18, 0]
[18, 0]]
gini = 0.0
samples = 3
value = [[3, 0]
[0, 3]
[3, 0]
[3, 0]]
VEC ≤ 6.354
gini = 0.089
samples = 13
value = [[3, 10]
[0, 13]
[13, 0]
[13, 0]]
gini = 0.0
samples = 1
value = [[0, 1]
[1, 0]
[1, 0]
[1, 0]]
gini = 0.0
samples = 2
value = [[0, 2]
[2, 0]
[2, 0]
[2, 0]]
gini = 0.0
samples = 7
value = [[0, 7]
[0, 7]
[7, 0]
[7, 0]]
J ≤ 3.204
gini = 0.125
samples = 6
value = [[3, 3]
[0, 6]
[6, 0]
[6, 0]]
gini = 0.0
samples = 3
value = [[3, 0]
[0, 3]
[3, 0]
[3, 0]]
gini = 0.0
samples = 3
value = [[0, 3]
[0, 3]
[3, 0]
[3, 0]]

References

Agrawal, Ankit, and Alok Choudhary. 2016. “Perspective: Materials Informatics and Big Data: Realization of the ‘Fourth Paradigm’ of Science in Materials Science.” *APL Materials* 4 (5):053208. https://doi.org/10.1063/1.4946894.

Allen, Todd, and Lizhen Tan. 2008. “Microstructural Effects on the Corrosion Behavior of Alloys and Ceramics.” National Association of Corrosion Engineers Houston TX. http://www.dtic.mil/docs/citations/ADA487962.

Allmann, R., and R. Hinek. 2007. “The Introduction of Structure Types into the Inorganic Crystal Structure Database ICSD.” *Acta Crystallographica Section A: Foundations of Crystallography* 63 (5):412–17. https://doi.org/10.1107/S0108767307038081.

Arganda-Carreras, Ignacio, Verena Kaynig, Curtis Rueden, Kevin W. Eliceiri, Johannes Schindelin, Albert Cardona, H. Sebastian Seung, and Robert Murphy. 2017. “Trainable Weka Segmentation: A Machine Learning Tool for Microscopy Pixel Classification.” *Bioinformatics* 33 (15):2424–26. https://doi.org/10.1093/bioinformatics/btx180.

ASM International. n.d. “ASM Micrograph Database™ - ASM International.” Accessed August 1, 2018. https://www.asminternational.org/materials-resources/online-databases/-/journal_content/56/10192/15468789/DATABASE.

Barrett, Charles A. 1992. “A Statistical Analysis of Elevated Temperature Gravimetric Cyclic Oxidation Data of 36 Ni- and Co-Base Superalloys Based on an Oxidation Attack Parameter.” https://ntrs.nasa.gov/search.jsp?R=19930008880.

———. 1997. “10,000-Hour Cyclic Oxidation Behavior at 982 C (1800 F) of 68 High-Temperature Co-, Fe-, and Ni-Base Alloys.” https://ntrs.nasa.gov/search.jsp?R=19970025059.

———. 2003. “A High Temperature Cyclic Oxidation Data Base for Selected Materials Tested at NASA Glenn Research Center.” NASA/TM-2003-212546. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20030093726.pdf.

Boutell, Matthew R., Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. “Learning Multi-Label Scene Classification.” *Pattern Recognition* 37 (9):1757–71. https://doi.org/10.1016/j.patcog.2004.03.009.

Boyer, Howard E., and Timothy L. Gall, eds. 1981. “Metallography: Measurement of Microstructure.” In *Metals Handbook Desk Edition*. American Society of Metals.

Brence, John R., and Donald E. Brown. 2002. “Data Mining Corrosion from Eddy Current Non-Destructive Tests.” *Computers & Industrial Engineering* 43 (4):821–40. https://doi.org/10.1016/S0360-8352(02)00142-0.

Brookhaven National Laboratory. n.d. “BNL | Center for Data-Driven Discovery (C3D).” Accessed August 3, 2018. https://www.bnl.gov/compsci/C3D/.

Brough, David B., Daniel Wheeler, and Surya R. Kalidindi. 2017. “Materials Knowledge Systems in Python—a Data Science Framework for Accelerated Development of Hierarchical Materials.” *Integrating Materials and Manufacturing Innovation* 6 (1):36–53. https://doi.org/10.1007/s40192-017-0089-0.

Curtarolo, Stefano, Wahyu Setyawan, Gus L. W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, et al. 2012. “AFLOW: An Automatic Framework for High-Throughput Materials Discovery.” *Computational Materials Science* 58 (June):218–26. https://doi.org/10.1016/j.commatsci.2012.02.005.

DMMI. 2014. “Big Data in Materials Research and Development: Summary of a Workshop.” Defense Materials Manufacturing and Infrastructure Standing Committee, National Research Council.

DoD. 2015. “Department of Defense Plan to Establish Public Access to the Results of Federally Funded Research.” Department of Defense, Washington D.C. http://www.dtic.mil/docs/citations/ADA614321.

Domínguez, L. Asensio, R. Goodall, and I. Todd. 2015. “Prediction and Validation of Quaternary High Entropy Alloys Using Statistical Approaches.” *Materials Science and Technology* 31 (10):1201–6. https://doi.org/10.1179/1743284715Y.0000000019.

Fernandez, Michael, Tom K. Woo, Christopher E. Wilmer, and Randall Q. Snurr. 2013. “Large-Scale Quantitative Structure–Property Relationship (QSPR) Analysis of Methane Storage in Metal–Organic Frameworks.” *The Journal of Physical Chemistry C* 117 (15):7681–89. https://doi.org/10.1021/jp4006422.

Field, Kevin G., Samuel A. Briggs, Philip D. Edmondson, Jack C. Haley, Richard H. Howard, Xunxiang Hu, Kenneth C. Littrell, Chad M. Parish, and Yukinori Yamamoto. 2016. “Database on Performance of Neutron Irradiated FeCrAl Alloys.” ORNL/TM-2016/335. Oak Ridge National Laboratory. https://info.ornl.gov/sites/publications/files/Pub68773.pdf.

Field, Kevin G., Samuel A. Briggs, Philip Edmondson, Xunxiang Hu, Kenneth C. Littrell, Richard Howard, Chad M. Parish, and Yukinori Yamamoto. 2015. “Evaluation on the Effect of Composition on Radiation Hardening and Embrittlement in Model FeCrAl Alloys.” ORNL/TM-2015/518. Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States). https://doi.org/10.2172/1253237.

Field, Kevin G., Samuel A. Briggs, Xunxiang Hu, Yukinori Yamamoto, Richard H. Howard, and Kumar Sridharan. 2017. “Heterogeneous Dislocation Loop Formation near Grain Boundaries in a Neutron-Irradiated Commercial FeCrAl Alloy.” *Journal of Nuclear Materials* 483 (January):54–61. https://doi.org/10.1016/j.jnucmat.2016.10.050.

Field, Kevin G., Samuel A. Briggs, Kumar Sridharan, Richard H. Howard, and Yukinori Yamamoto. 2017. “Mechanical Properties of Neutron-Irradiated Model and Commercial FeCrAl Alloys.” *Journal of Nuclear Materials* 489 (June):118–28. https://doi.org/10.1016/j.jnucmat.2017.03.038.

Georgia Tech. n.d. “MATIN - Home.” Accessed August 2, 2018. https://matin.gatech.edu/.

Gild, Joshua, Yuanyao Zhang, Tyler Harrington, Sicong Jiang, Tao Hu, Matthew C. Quinn, William M. Mellor, Naixie Zhou, Kenneth Vecchio, and Jian Luo. 2016. “High-Entropy Metal Diborides: A New Class of High-Entropy Materials and a New Type of Ultrahigh

Temperature Ceramics." *Scientific Reports* 6 (November):37946. https://doi.org/10.1038/srep37946.

Gorai, Prashun, Duanfeng Gao, Brenden Ortiz, Sam Miller, Scott A. Barnett, Thomas Mason, Qin Lv, Vladan Stevanović, and Eric S. Toberer. 2016. "TE Design Lab: A Virtual Laboratory for Thermoelectric Material Design." *Computational Materials Science* 112 (February):368–76. https://doi.org/10.1016/j.commatsci.2015.11.006.

Groeber, Michael A., and Michael A. Jackson. 2014. "DREAM.3D: A Digital Representation Environment for the Analysis of Microstructure in 3D." *Integrating Materials and Manufacturing Innovation* 3 (1):5. https://doi.org/10.1186/2193-9772-3-5.

Guo, Sheng. 2015. "Phase Selection Rules for Cast High Entropy Alloys: An Overview." *Materials Science and Technology* 31 (10):1223–30. https://doi.org/10.1179/1743284715Y.0000000018.

Guo, Sheng, Chun Ng, Jian Lu, and C. T. Liu. 2011. "Effect of Valence Electron Concentration on Stability of Fcc or Bcc Phase in High Entropy Alloys." *Journal of Applied Physics* 109 (10):103505. https://doi.org/10.1063/1.3587228.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer.

Hautier, Geoffroy, Anna Miglio, Gerbrand Ceder, Gian-Marco Rignanese, and Xavier Gonze. 2013. "Identification and Design Principles of Low Hole Effective Mass *p*-Type Transparent Conducting Oxides." *Nature Communications* 4 (August):2292. https://doi.org/10.1038/ncomms3292.

IAEA. 2009. "Integrity of Reactor Pressure Vessels in Nuclear Power Plants: Assessment of Irradiation Embrittlement Effects in Reactor Pressure Vessel Steels." NP-T-3.11. International Atomic Energy Agency.

Information International Associates. 2012. "E-Ceramics 2012: The Availabilty, Quality, and Location of Ceramic Property Data."

Jackson, A. G., Z. Pawlak, and S. R. LeClair. 1998. "Rough Sets Applied to the Discovery of Materials Knowledge - ScienceDirect." *Journal of Alloys and Compounds* 279:14–21.

Jain, Anubhav, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, et al. 2013. "Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation." *APL Materials* 1 (1):011002. https://doi.org/10.1063/1.4812323.

Johnson, Sherena G. 2010. "Materials." NIST. August 27, 2010. https://www.nist.gov/srd/search-srd-topic/materials.

Joint Research Centre. n.d. "MatDB." Accessed May 10, 2018. https://odin.jrc.ec.europa.eu/alcor/Main.jsp.

Kalidindi, Surya R., and Marc De Graef. 2015. "Materials Data Science: Current Status and Future Outlook." *Annual Review of Materials Research* 45 (1):171–93. https://doi.org/10.1146/annurev-matsci-070214-020844.

Kalidindi, Surya R., Stephen R. Niezgoda, and Ayman A. Salem. 2011. “Microstructure Informatics Using Higher-Order Statistics and Efficient Data-Mining Protocols.” *JOM* 63 (4):34–41. https://doi.org/10.1007/s11837-011-0057-7.

Kong, Chang Sun, Wei Luo, Sergiu Arapan, Pierre Villars, Shuichi Iwata, Rajeev Ahuja, and Krishna Rajan. 2012. “Information-Theoretic Approach for the Discovery of Design Rules for Crystal Chemistry.” *Journal of Chemical Information and Modeling* 52 (7):1812–20. https://doi.org/10.1021/ci200628z.

Lupulescu, Afina, Tyler Flowers, Linda Vermillion, and Scott Henry. 2015. “ASM Micrograph Database™.” *Metallography, Microstructure, and Analysis* 4 (4):322–27. https://doi.org/10.1007/s13632-015-0212-6.

Mao, Huahai, Hai-Lin Chen, and Qing Chen. 2017. “TCHEA1: A Thermodynamic Database Not Limited for ‘High Entropy’ Alloys.” *Journal of Phase Equilibria and Diffusion* 38 (4):353–68. https://doi.org/10.1007/s11669-017-0570-7.

Metal AM. 2018. “LLNL Investigates Machine Learning for Defect Prevention in Metal Additive Manufacturing.” *Metal Additive Manufacturing* (blog). September 17, 2018. http://www.metal-am.com/llnl-investigates-machine-learning-for-defect-prevention-in-metal-additive-manufacturing/.

Miracle, D. B. 2015. “Critical Assessment 14: High Entropy Alloys and Their Development as Structural Materials.” *Materials Science and Technology* 31 (10):1142–47. https://doi.org/10.1179/1743284714Y.0000000749.

———. 2018. “Hi Entropy Alloys (CCAs): A New Approach for Alloy Discovery and Development.” presented at the DOD HEA Workshop, Institute for Defense Analyses, May 3.

M&MP COI. 2017. “Joint Program Plan, Department of Defense Reliance 21.” Materials and Manufacturing Processes Community of Interest.

Morefield, Sean, Susan A. Drozdz, Vincent F. Hock, William H. Abbott, David Paul, and Jana L. Jackson. 2009. “Development of a Predictive Corrosion Model Using Locality-Specific Corrosion Indices.” ERDC/CERL-TR-09-22. Engineer Research and Development Center Champaign IL Construction Engineering Research Lab. http://www.dtic.mil/docs/citations/ADA528275.

Motta, Arthur T., Adrien Couet, and Robert J. Comstock. 2015. “Corrosion of Zirconium Alloys Used for Nuclear Fuel Cladding.” *Annual Review of Materials Research* 45 (1):311–43. https://doi.org/10.1146/annurev-matsci-070214-020951.

Niezgoda, Stephen R., Anand K. Kanjarla, and Surya R. Kalidindi. 2013. “Novel Microstructure Quantification Framework for Databasing, Visualization, and Analysis of Microstructure Data.” *Integrating Materials and Manufacturing Innovation* 2 (1):3. https://doi.org/10.1186/2193-9772-2-3.

NIST. n.d. “Materials Resource Registry Curation System.” Accessed August 3, 2018. https://materials.registry.nist.gov/.

Over, H.H., E. Wolfart, W. Dietz, and L. Toth. 2005. “Mat-DB: A Web-Enabled Materials Database to Support European R&D Projects and Network Activities.” *Advanced Engineering Materials* 7 (8):766–70. https://doi.org/10.1002/adem.200500059.

PEO Land Systems. 2017. “Advanced Technology Investment Plan.” US Marine Corps.

PRISMS. n.d. “PRISMS Home.” Accessed June 4, 2018. http://www.prisms-center.org/#/home.

Qu, Xiaohui, Anubhav Jain, Nav Nidhi Rajput, Lei Cheng, Yong Zhang, Shyue Ping Ong, Miriam Brafman, Edward Maginn, Larry A. Curtiss, and Kristin A. Persson. 2015. “The Electrolyte Genome Project: A Big Data Approach in Battery Materials Discovery.” *Computational Materials Science* 103 (June):56–67. https://doi.org/10.1016/j.commatsci.2015.02.050.

QuesTek Innovations LLC. 2018. “High Entropy Alloys.” 2018. https://www.questek.com/high-entropy-alloys.html.

Rajan, Krishna. 2015. “Materials Informatics: The Materials ‘Gene’ and Big Data.” *Annual Review of Materials Research* 45 (1):153–69. https://doi.org/10.1146/annurev-matsci-070214-021132.

Ramprasad, Rampi, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. 2017. “Machine Learning in Materials Informatics: Recent Applications and Prospects.” *Npj Computational Materials* 3 (1):54. https://doi.org/10.1038/s41524-017-0056-5.

RE3Data. n.d. “Registry of Research Data Repositories.” Accessed August 3, 2018. https://www.re3data.org/.

Rost, Christina M., Sachet, Edward, Borman, Trent, Moballegh, Ali, Dickey, Elizabeth C., Hou, Dong, Jones, Jacob L., Curtarolo, Stefano, and Maria, Jon-Paul. 2015. “Entropy-Stabilized Oxides.” *Nature Communications* 6:8485.

Russ, John C. 2011. *The Image Processing Handbook*. 6th ed. Boca Raton: CRC Press.

Saal, James E., Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. 2013. “Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD).” *JOM* 2013:1501–9.

Sanvito, Stefano, Corey Oses, Junkai Xue, Anurag Tiwari, Mario Zic, Thomas Archer, Pelin Tozman, Munuswamy Venkatesan, Michael Coey, and Stefano Curtarolo. 2017. “Accelerated Discovery of New Magnets in the Heusler Alloy Family.” *Science Advances* 3 (4):e1602241. https://doi.org/10.1126/sciadv.1602241.

Smialek, J. L., and C. A. Barrett. 1997. “Design for Oxidation Resistance.Pdf.” In *ASM Handbook: Materials Selection and Design*. Vol. 20. ASM International.

Smialek, J. L., and P. J. Bonacuse. 2016. “Compositional Effects on the Cyclic Oxidation Resistance of Conventional Superalloys.” *Materials at High Temperatures*.

Smialek, J. L., J. A. Nesbitt, C. A. Barrett, and C. E. Lowell. 2000. “Cyclic Oxidation Testing and Modelling: A NASA Lewis Perspective.” https://ntrs.nasa.gov/search.jsp?R=20000033849.

Sziroczak, D., and H. Smith. 2016. “A Review of Design Issues Specific to Hypersonic Flight Vehicles.” *Progress in Aerospace Sciences* 84 (July):1–28. https://doi.org/10.1016/j.paerosci.2016.04.001.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. “Data Sharing by Scientists: Practices and Perceptions.” *PLOS ONE* 6 (6):e21101. https://doi.org/10.1371/journal.pone.0021101.

Troparevsky, M. Claudia, James R. Morris, Paul R. C. Kent, Andrew R. Lupini, and G. Malcolm Stocks. 2015. “Criteria for Predicting the Formation of Single-Phase High-Entropy Alloys.” *Physical Review X* 5 (1):011041. https://doi.org/10.1103/PhysRevX.5.011041.

U.S. Naval Research Laboratory. 2018. “Center for Corrosion Science & Engineering (6130).” 2018. https://www.nrl.navy.mil/chemistry/research/6130.

Walt, Stéfan van der, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. 2014. “Scikit-Image: Image Processing in Python.” *PeerJ* 2 (June). https://doi.org/10.7717/peerj.453.

Wang, Jy-An John. 2010. “Lessons Learned From Developing Reactor Pressure Vessel Steel Embrittlement Database.” ORNL/TM-2010/20. Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States); High Temperature Materials Laboratory. https://doi.org/10.2172/985285.

Wang, Jy-An John, and Ranjit Subramani. 2008. “PR-EDB: Power Reactor Embrittlement Database Version 3.” ORNL/TM--2006/605. Oak Ridge National Laboratory (United States). Funding organisation: ORNL work for others (United States). https://doi.org/10.2172/932609.

Xu, Yibin, Masayoshi Yamazaki, and Pierre Villars. 2011. “Inorganic Materials Database for Exploring the Nature of Material.” *Journal of Applied Physics* 50:11RH02.

Yang, X., and Y. Zhang. 2012. “Prediction of High-Entropy Stabilized Solid-Solution in Multi-Component Alloys.” *Materials Chemistry and Physics* 132 (2):233–38. https://doi.org/10.1016/j.matchemphys.2011.11.021.

Zhang, Yong, Ting Ting Zuo, Zhi Tang, Michael C. Gao, Karin A. Dahmen, Peter K. Liaw, and Zhao Ping Lu. 2014. “Microstructures and Properties of High-Entropy Alloys.” *Progress in Materials Science* 61 (April):1–93. https://doi.org/10.1016/j.pmatsci.2013.10.001.

Zhou, Jieyang, Zhang, Jinyong, Zhang, Fan, Niu, Bo, Lei, Liwen, and Wang, Weimin. 2018. “High-Entropy Carbide: A Novel Class of Multicomponent Ceramics.” *Ceramics International*, August. https://doi.org/10.1016/j.ceramint.2018.08.100.

Zinkle, S. J., and L. L. Snead. 2014. “Designing Radiation Resistance in Materials for Fusion Energy.” *Annual Review of Materials Research* 44 (1):241–67. https://doi.org/10.1146/annurev-matsci-070813-113627.

Abbreviations

AIC	Akaike Information Criterion
ASM	American Society for Metals
BCC	Body Centered Cubic
CCA	Complex Concentrated Alloy
DoD	Department of Defense
DOE	Department of Energy
DTIC	Defense Technical Information Center
FCC	Face Centered Cubic
HEA	High Entropy Alloy
IAEA	International Atomic Energy Agency
ICSD	Inorganic Crystal Structure Database
LARS	Least Angle Regression
MATIN	Materials Innovation Network
NASA	National Aeronautics and Space Administration
NIST	National Institute of Standards and Technology
ORNL	Oak Ridge National Lab
PCA	Principal Component Analysis
PC	Principal Component
PSP	Processing-structure-property
QA	Quality Assurance
RBF	Radial Basis Function
RPV	Reactor Pressure Vessel
SVM	Support Vector Machine
TEM	Transmission Electron Microscope
TPS	Thermal Protection System
VEC	Valence Electron Concentration

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From–To)*
September 2018	Draft Final	JAN 2018 – AUG 2018

4. TITLE AND SUBTITLE

Data Science for Materials in Extreme Environments

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Swallow, Jessica G.

5d. PROJECT NUMBER
CRP 2229

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-9301

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

10. SPONSOR/MONITOR'S ACRONYM(S)

IDA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (20 November 2018).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report examines the value of statistical learning and data mining approaches to the design, development, and modeling of materials in extreme environments, which is a cross-cutting problem within the DoD. Existing tools and databases are briefly surveyed, along with challenges implementing statistical learning methods in the DoD materials research and testing communities. Multi-label classification and microstructural variable quantification are highlighted as important techniques for understanding materials in extreme environments and developing useful processing-structure-property linkages. Case studies exploring classification of high entropy alloy structure, prediction of radiation-induced embrittlement, and computational image processing to characterize material microstructure are presented. The need to document and standardize image analysis methods is emphasized. Barriers to enhancing modeling and design of materials for extreme environments using statistical learning approaches remain, including limited data availability, standardization, and analysis tools, and limited workforce proficiency in techniques for data storage, curation, and analysis.

15. SUBJECT TERMS

computational image processing; computationally-generated data; extreme environments; fuzzy logic; materials data science tools; materials science; multi-label classification; statistical learning tools

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Buckley, Leonard J.
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	68	19b. TELEPHONE NUMBER *(include area code)* 703-578-2800

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18