DATA ANALYTICS APPROACH FOR STUDYING STRUCTURE-PROPERTY RELATIONSHIPS IN GRADIENT ALUMINUM COMPOSITE

by

AMIT KUMAR

Submitted in partial fulfillment of the requirements

For the degree of Masters of Science

Thesis Adviser: Dr. Jennifer L W Carter

Department of Materials Science and Engineering

CASE WESTERN RESERVE UNIVERSITY

May, 2016
Data Analytics Approach for studying Structure-Property
Relationships in Gradient Aluminum Composite

Case Western Reserve University
Case School of Graduate Studies

We hereby approve the thesis\(^1\) of

AMIT KUMAR

for the degree of

Masters of Science

\(^1\)We certify that written approval has been obtained for any proprietary material contained therein.

Dated: 03/16/2016
## Table of Contents

List of Tables ................................................................. v

List of Figures .............................................................. vi

Acknowledgements ........................................................... ix

Abstract ........................................................................... x

Chapter 1. Introduction ...................................................... 1

Chapter 2. Literature Review ............................................. 4

  What is Data Science? ......................................................... 4

  \( < P_r | M | P_e > \) Interdependence ........................................ 5

  Neural Networks .............................................................. 7

  Recent Push: ICME .......................................................... 10

  New Network Models ....................................................... 12

Chapter 3. Methods and Data Collection .......................... 13

  Statistical Aspects of Data Collection ................................. 14

  Material ........................................................................ 14

  Quantification Techniques .................................................. 19

  Experimental Specifications .............................................. 24

Chapter 4. Exploratory Data Analysis and Semi-gSEM Performance Modeling 29

  Basic Steps in EDA .......................................................... 29

  Semi-gSEM ....................................................................... 34

  Model - Training/Testing/Validation ..................................... 37
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Conditions needed for Quantitative XEDS (Reproduced from Ritchie\textsuperscript{74,75}).</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of different techniques for chemical analysis. “*” means values are approximate. Actual value might change depending on the material and energy of incident beam.</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>Uni-variate functionals forms evaluated by semi-gSEM.</td>
<td>35</td>
</tr>
<tr>
<td>C.1</td>
<td>Uni-variate functionals forms evaluated by semi-gSEM (v0.4.5)</td>
<td>58</td>
</tr>
<tr>
<td>C.2</td>
<td>Uni-variate functionals forms evaluated by semi-gSEM strength branch</td>
<td>59</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Data-analytics schematic illustrating that mapping functions $\Phi$ and $\Psi$ between material measures that can be developed to predict and design alloy performance. 6

2.2 Hall Petch Relationship $(\sigma = \sigma_0 + k_y/\sqrt{d})$. High $k_y$ change grain size to affect yield, ductility and toughness. Low $k_y$ change grain size to affect ductility, toughness, and not strength. Reproduced from Meyers and Chawla$^{58}$. 6

2.3 Ashby plot of fracture toughness and modulus for metals, alloys, ceramic, glasses, polymers and metallic glasses. The contours indicates constant ratios of fracture toughness and elastic modulus. (Reproduced from Ashby and Greer$^9$) 8

2.4 Integrating databases and computational materials science tools. DFT, density functional theory; MD, molecular dynamics; TTT, time temperature transformation curve; and CCT, continuous cooling transformation curves. (Reproduced from ICME Report$^{38}$) 11

3.1 Phase Diagram of Al-Zn$^{66}$ with range of interest highlighted with a Black Bar (3 - 4.5 at% Zn). 15

3.2 Phase Diagram of Mg-Zn$^{65}$. 16

3.3 Phase Diagram of Al-Mg$^{67}$ with range of interest highlighted with a Black Bar (3.5 - 5 at% Mg). 17

3.4 Schematic of the samples with respective directions. 25
3.5 Box Plot of the variation in the hardness measurements at different loads as calculated from 22 measurements at each loading condition.

3.6 Schematic showing how the data is collected for training dataset. Left bottom (soft side or 5xxx surface) was taken as a reference point. Microhardness measurements were done at an interval of 2.54mm in both directions, and XEDS measurements at the neighboring positions (twice at each location).

4.1 Pairs-wise plot and linear correlation coefficient between atomic percentage of elements (Al-Mg-Zn-Cu) in the composite and Hardness for the training data set. X and Z represent the local coordinate system as defined in Figure 3.4. r and p values in the upper right quadrant are linear correlation coefficients.

4.2 Manual assessment for Hardness and Mg concentration.

4.3 Simplified semi-gSEM Network Model

4.4 Graphical representation showing high variability in data.

4.5 Residual Plots of functional form - \( H = 0.24e^{Zn} - 29\sqrt{Mg} + 247 \).

4.6 Pairs-wise plot and linear correlation coefficient between atomic percentage of (Al-Mg-Zn-Cu) in the composite and Hardness for the test data set. X and Z represents the same local co-ordinate system as defined for Figure 4.1. r and p values in the upper right quadrant are linear correlation coefficients.

4.7 Predicted hardness values (red) based on measured compositional gradient in comparison to measured hardness values. (Top) Overlapped with z-axis. (Bottom) Scatter in predicted values.
B.1 Chi-square Distribution Table. 56

C.1 Table generated of object of class sgSEM, before and after the addition of new functional forms. 60

C.2 Summary of the object of class sgSEM, before and after the addition of new functional forms. 61

C.3 Semi-gSEM network model plots generated from object of class sgSEM using DiagrammeR package. 62
Acknowledgements

I would like to thank all of the people at Case Western Reserve who made this work possible. My advisor, Prof. Jennifer L. W. Carter, for her guidance and support during my time at Case, Prof. Roger H. French, and everyone else who is part of Solar Durability and Lifetime Extension (SDLE) Center, for constantly tweaking my code chunks, Prof. John J. Lewandowski, and members of Advanced Manufacturing and Mechanical Reliability Center (AMMRC), for giving me feedback at various stages of this study, especially, for experimental designs.

Also, members of Swagelok Center for Surface Analysis of Materials (SCSAM), with whom I got acquainted to electron microscopy, and was successful in acquiring large amounts of data for my study.

This research was funded by the Army Research Office Short Term Innovative Research (STIR) program: W911NF-14-0549. Material was supplied by the Alcoa Alloy Technology Division. I gratefully acknowledge their support. Also, I would like to thank researchers in the Statistical Research Computing and Collaboration (SR2c) for their input on the study design for this work.
Abstract

Data Analytics Approach for studying Structure-Property Relationships in Gradient Aluminum Composite

Abstract

by

AMIT KUMAR

Architecturally graded composites can enable the production of materials whereby conflicting design constraints are reconciled for different locations in the material (i.e., strength in one region, with ductility in another). The optimization of these materials requires a priori knowledge of how different measures of the architecture (i.e., composition, microstructure, geometry) are interdependent and combine to control the material's performance. In this work, a data-analytic approach relating composition, microstructure, and properties was applied to predict the mechanical performance of an architecturally graded aluminum composite with a gradient in composition between Al-5456 and Al-7055. Predictions were made using a semi-supervised, generalized, structural equation modeling (semi-gSEM) approach. This network modeling approach allows for the statistical validation of diagnostic models of material performance. This work successfully developed a network model that can predict material performance of a gradient aluminum composite with statistical significance of 0.6 adj-$R^2$. 
1 Introduction

The fine-tuning of properties during the materials design process is essentially a multi-criteria, multi-dimensional, optimization, in which secondary aspects, such as the availability of alloy components, and the achievable microstructures are optimized for the desired alloy performance. Quantifying the contributors to overall material performance is a complex problem, as multiple mechanisms can act in concert and often individual mechanisms are not fully understood\(^4,21,59,69\). The reliance upon the physics of mesostructure evolution, makes predicting performance computationally demanding, as these structures often span multiple length scales, where parameters influencing and the governing physics changes with the scale. Integrated Computational Materials Engineering (ICME\(^38\)) has taken up this task by integrating existing physics-based models at different length scales.

For purely predictive models, neural networks\(^{14,23}\) (i.e., regression analysis) can quickly produce models of complex systems where few of the physical mechanisms are known. Neural networks are fairly simple to execute on complex systems where physics-based understanding is lacking\(^{14}\), and they provide design insight when interpolating in the processing and mesostructure evolution landscape. There has been great engineering
design success in applying this technique to steel\textsuperscript{70,76}, nickel\textsuperscript{12,55}, and titanium alloys\textsuperscript{56}. Unfortunately, these models tend not to provide insights into the physics.

Domain-Guided Statistical Design provides a new method of designing optimal manufacturing and materials with additional physics insight that is not possible with neural networks, and with reduced complexity as compared to ICME approaches\textsuperscript{45}. Conventional statistical design methods include structural equation modeling, a multi-variate regression approach that generates a system of equations to identify relationships and latent variables from observed dependent (endogenous) and independent (exogenous) variables. Typical structural equation models only fit a system of linear equations\textsuperscript{5}, and are independent of domain information; but the methodologies have been updated recently to allow for higher-order systems of equations to be considered based on pre-existing domain knowledge. This type of analysis has been termed semi-supervised generalized system equation modeling (semi-gSEM); semi-supervised because we can incorporate pre-existing domain knowledge and generalized because additional functional forms besides linear fits are accommodated. Semi-gSEM has been successfully applied to explore the often competing degradation mechanisms in acrylic polymers\textsuperscript{61}, and layered photovoltaic materials and systems\textsuperscript{17,18,40,61}.

In this work, a domain-guided diagnostic model had been developed using semi-gSEM, which integrates existing physics and provides additional insights. The following chapter will give a more detailed description of various data-driven models and their predictive capabilities, while introducing data science. The third chapter, contains specific details of experiments, e.g., material description, relevant domain study, basic outline of data collection through different techniques, and specific details of experiments. The fourth chapter, systematically explains the steps involved in analysis, with
tools used for building quantified predictive model. The fourth chapter introduces a hierarchical methodology which can be applied and modulated for other problems in materials science, with strong discussions on its applicability. Finally, discussions and conclusions, which can be drawn from the study.
2 Literature Review

2.1 What is Data Science?

Data science is an interdisciplinary field focused on the processes and systems to extract knowledge and insights from data\textsuperscript{1,27,33}. From a materials science perspective, it is different from physics-based multiscale materials modelling which involves solving the governing field equations under specified materials constitutive laws by imposing boundary and initial conditions. Data science techniques extract embedded low-dimensional linkages between the various inputs and outputs involved in the numerical simulation.

Primarily, data science is composed of two components\textsuperscript{50}, i) data management (informatics), which includes robust and reliable storage, aggregation, and sharing; and ii) data analytics, aimed at mining the embedded high-value information via statistical analyses, dimensionality reduction, pattern recognition, regression analyses, and machine learning. This first step is critical and necessary, as it includes not just the files produced by the instrument, but also various other pertinent details about the experiment referred as metadata (e.g., environmental conditions, details of how the test specimens were prepared, instrument settings, etc.). The central purpose of metadata is to enhance the utility of the data by providing important information on the context and
content, thereby allowing the data to be re-discovered in appropriate searches by potential future users. The second task revolves around data analytics, or building data-driven models (DDM) without any reference to the underlying physics of the process. In situations where the underlying scientific principles are perhaps not known or either too complex to capture, these models can be helpful. But, where the underlying physics does exist, it needs to be sufficiently rigorous so any physical trend that is significant for a particular process should be reflected in the data itself and the model should be able to capture it and predict the correct physical trends.

2.2 $< P_r|M|P_e >$ Interdependence

Whenever a material is being created, developed, or produced, the properties or phenomena the material exhibits are of central concern. Established science shows that the properties and phenomena associated with a material are intimately related to its composition and structure at all levels, including which atoms are present and how they are arranged, and that this structure is the result of synthesis and processing. Finding quantitative $< $ Processing $| $ Microstructure $| $ Performance $> ($ $< P_r|M|P_e > $) correlations or mapping functions ($\Phi$ and $\Psi$), allow us to design/predict the products/performance (Figure 2.1).

For example, the Hall-Petch (H-P) relationship\textsuperscript{20,77} (Figure 2.2), is a functional map from the microstructure measure of grain size ($d$) to the performance measure of strength ($\sigma$) such that the mapping function, $\Psi(\sigma, d)$, is $\sigma = \sigma_0 + k_y / \sqrt{d} + \epsilon$, where $\sigma_0$ and $k_y$ are experimentally validated constants and $\epsilon$ is a measure of the fitting error. H-P relation is
Literature Review

Figure 2.1. Data-analytics schematic illustrating that mapping functions $\Phi$ and $\Psi$ between material measures that can be developed to predict and design alloy performance.

Figure 2.2. Hall Petch Relationship ($\sigma = \sigma_0 + k_y/\sqrt{d}$). High $k_y$ change grain size to affect yield, ductility and toughness. Low $k_y$ change grain size to affect ductility, toughness, and not strength. Reproduced from Meyers and Chawla.

often used as a guide for materials processing to obtain more desired properties. Correlating the average grain size to the various thermo-mechanical deformation cycles, has
also provided new insights to further improve the performance\textsuperscript{24,37,49}. This, however, also leads to the belief that the average grain size strongly influence the overall mechanical properties of the metal product in service\textsuperscript{35,52,60}, although it is not the only factor influencing the final performance. This relation works, in spite of not having a totally accepted theory/model.

Ashby plots\textsuperscript{8}(e.g. Figure 2.3) added one more dimension, and showed that by merging phenomenological relationships (or $< P_r|M|P_e>$ relationships) in materials properties with discrete data on specific material characteristics, one can begin to develop patterns for classification of materials behavior. This step, also provided a methodology to establish common $< P_r|M|P_e>$ relationships across seemingly different classes of materials. This section highlighted that some early examples of $< P_r|M|P_e>$ interdependence can also be seen as a certain class of DDM, which are empirical and diagnostic in nature.

\section{2.3 Neural Networks}

Neural networks are another class of DDM that have been applied to materials science\textsuperscript{14}. The study of neural networks highlights many aspects of DDM and the complexity that exists in materials science databases.

In a conventional feedforward neural network, information flows from the input to the output layer via a hidden layer in between. The nodes in the hidden layer receive information from the input layer, where each connection is attributed its individual weight. This information is aggregated and processed in the hidden layer, usually through some preset nonlinear functions such as a hyperbolic tangent\textsuperscript{14} (due to
Figure 2.3. Ashby plot of fracture toughness and modulus for metals, alloys, ceramic, glasses, polymers and metallic glasses. The contours indicates constant ratios of fracture toughness and elastic modulus. (Reproduced from Ashby and Greer\(^9\))

its flexibility). The information is then passed on to the output layer using another set of weights. The optimized values of the weights are obtained by solving a minimizing problem of the errors in the outputs generated from training data set inputs. One hidden layer may not be sufficient in some cases, therefore, multiple hidden layers can be introduced to capture highly complex non-linear relationships.

Neural networks do not produce a diagnostic model, as the nature of interactions which it captures are hidden in the values of weights, and these weights are not easy to interpret. For example, there may exist more than just pairwise interactions, in which case it becomes difficult to examine the weights. This limits its applicability, as it cannot
provide any insights into the physics of the process, and produces only empirical models. The most common method to use the neural network models is to make predictions and see how well they perform by comparing the difference between the actual and the predicted value (or residuals).

Industrial data in materials science is routinely associated with a very significant amount of both random and systematic noise, and a good DDM needs to tackle that problem efficiently. Noise present in a systematic manner is often easily detectable and easy to filter out. The problem, however, is far more severe for random noise attributed to diverse sources and various reasons. One such scenario is an attempt to capture all the apparent trends in a data set, which eventually leads to the overfitting problem\textsuperscript{14,28}. Another extreme scenario would be the underfitting\textsuperscript{14,28} of data, where the model is devoid of any noise but is unable to capture the physical trends. The complexity of a model may be expressed in terms of the number of parameters (or predictors) it uses. The more complex the model becomes, the greater the chances of overfitting, and conversely, at a very low complexity underfitting may readily set in. The task is therefore to come up with a model of the right complexity that will neither over- nor underfit. Often, to avoid these difficulties, the experimental data is divided into two sets, a training dataset and a test dataset. The model is produced using only the training data, then the test data are subsequently used to check the validity of model predictions from previously unseen data.

Neural networks are clearly extremely useful in recognising patterns in complex data (e.g., Bhadeshia et al.\textsuperscript{15}, Fujii et al.\textsuperscript{41}), but often, models tends to overfit or underfit, if one is not cautious about these affects. Also, the empirical nature of neural network models hinders its assimilation with the well-established concepts of materials science.
2.4 Recent Push: ICME

ICME\textsuperscript{38} is an extended version of miniature Ashby plots. ICME spans across multiple length scales, integrating already established models at each length scale, from processing to properties. This approach calls for a single, central source, in which all relevant data is captured and consolidated from birth to death and a variety of software tools for full integration\textsuperscript{6}. These tools are diverse and range from the atomic level to the continuum level and from thermodynamic models to science-based property models. In Arnold et al.\textsuperscript{7}, an effort towards establishing such a database was discussed, using micromechanics analysis toolset for integration. However, ICME is not limited to micromechanics. The technique used for integration depends on the length scale and the material system.

For example, in micromechanics, the individual materials (typically referred to as constituents or phases) that make up a composite are each treated as continuous via continuum mechanics models, with their individual properties and arrangement dictating the overall behavior of the composite material. This works both ways, continuity conditions are used to formulate a set of semi-analytical linear algebraic equations that are solved for the local strains in subcells in terms of globally applied strains or stresses, then, local constitutive laws can be utilized to obtain the local stresses in subcell, and vice versa\textsuperscript{3}.

Micromechanics, is one of the many data integration techniques used in ICME, but it is still able to capture the principle behind ICME. ICME integration methods uses various length scale-specific relationships across the scales, from processing all the way to product performance (Figure 2.4). They are computationally demanding and require physics-based multi-scale modeling with issues of parameter passing between scales
to develop predictive models. This methodology also makes ICME design models, not generalizable across multiple material classes.
2.5 New Network Models

Network models are a broad category of statistical models for studying multi-variate relationships in complex systems and have been developed and applied in many fields such as genomics/proteomics, climate modeling and materials. Network modeling has benefited from advances in data-science, analytics, and distributed computing which have expanded the size of datasets and the scale of pathway networks that can be analyzed. In addition, advances in graph-theory approaches have aided the visualization of the multiplicity of relationships between graph nodes and pathways that form the variable relationships.

In this study, structural equation modeling (SEM), a network modeling approach used extensively in the social sciences, has been applied and extended to these types of models because they can capture statistically significant relationships amongst endogenous and exogenous variables, and provide insights to identify latent (unmeasured or hidden) variables. These extended SEM models, termed semi-supervised, generalizable SEM (semi-gSEM), are constrained by the functional forms guided from domain science, and often do not produce predictive models as accurate as a fitted neural network model. The semi-gSEM are both predictive (similar to neural network models) and also diagnostic/inference model. This diagnostic capability allows internal network and pathway relationships among the variables to provide insight on the physics of the system.
3 Methods and Data Collection

One may naturally assume that having large amounts of data is critical for any serious data-analytics studies. What constitutes “enough” data in materials science applications, however, can vary significantly. In studying structural ceramics, for instance, fracture toughness measurements \( K_{IC} \) (depends on Weibull stats, and what controls \( K_{IC} \)) are difficult to make and in some of the more complex materials just a few careful measurements can be of great value. Similarly, having reliable measurements on fundamental constants or properties, such as elastic modulus \( E \) or band gap energy \( E_g \), for a given material involves very detailed measurement and/or computational techniques.

In this chapter, statistical aspects of data collection are discussed, in general, for Network Models like SEM, and particularly for techniques used in data collection. Also, this chapter addresses material description, with its domain study and a discussion on “which technique is best for quantifying the compositional gradient in light of statistical analysis” and microhardness indentation, its formulations, and a small statistical study to check the variability in measurements by varying different loads. In the end, the chapter discusses experiment specifics, which give detailed description of how the experiments are conducted and collected data was ingested for analysis.
3.1 Statistical Aspects of Data Collection

For Network Models, datasets collected should be statistically significant for training, testing, and validation of the model. For statistical significant datasets (which means that the relationships between variables is caused by something other than random chance), there should be enough variability that the “p-value” (representing the probability that the random chance could explain the result) is typically less than 5%. Estimating the number of observations for p-values “less than or equal to 5%” can be challenging. But a fair trade off is to have a minimum of 6 independent observations of each variable (i.e., strength, grain size) that can capture the nature of the relationship, and enable differentiating between a linear, quadratic, exponential, and other fits between the variables. If it is still difficult to differentiate different functional fits or the p-value is more than 5%, increasing the number of observations is necessary.

3.2 Material

This work focused on the development of mapping functions and pathway relationships between \(< P_r | M | P_e >\) variables. An architecturally graded aluminum composite produced by co-casting was used to validate the mapping function technique for structural materials. This material has a continuous gradient in composition through the plate thickness (short transverse) from a stronger, precipitate strengthened, aluminum alloy (7055) to a softer, solid solution strengthened, aluminum alloy (5456). This diffuse gradient provides an opportunity for multiple independent observations of microstructural and performance metrics in a single material by which a statistical model can be derived with minimal material volume required.
Figure 3.1. Phase Diagram of Al-Zn with range of interest highlighted with a Black Bar (3 - 4.5 at% Zn).

3.2.1 Domain Study

The alloying additions of Zn and Mg are primarily responsible for the strengthening mechanisms in 7xxx and 5xxx series aluminum alloys respectively. The phase diagrams provide insights into how these alloying additions provide strengthening mechanisms. The Al-Zn phase diagram (Figure 3.1) indicates that there is limited solubility of Zn in Al at room temperature resulting in the precipitation of the Zn phase. The Zn-Mg phase diagram (Figure 3.2) indicates that Mg is insoluble in Zn therefore it can be assumed that Zn precipitates out of Al solution with little to no interaction with Mg.
Zn Precipitate in Al-Mg Matrix (or Precipitate Strengthening). Strength depends on precipitate size and shape, volume fraction, nature of boundary between precipitate and matrix, and modulus (whether it can deform or not, responsible only in elastic regime) in comparison with the matrix.

Orowan (in 1948\textsuperscript{57}) gave the theory of strengthening of alloys by non-deformable particles. Since then it has been refined by various authors\textsuperscript{44}, but it serves as the basis for the discussion. Theoretically:

\[
\sigma_{Or} = M \frac{0.4Gb}{\pi(1-\nu)^{1/2}} \frac{ln(d/b)}{\lambda}
\]
Figure 3.3. Phase Diagram of Al-Mg$^{67}$ with range of interest highlighted with a Black Bar (3.5 - 5 at% Mg).

where $M$ is the Taylor factor (a metric of crystallographic orientation), $G$ is the shear modulus of aluminum, $b$ is the Burgers vector, $\nu$ is the Poisson's ratio, and $\bar{d} = \sqrt{2/3}d$ and $\bar{\lambda} = \bar{d}(\sqrt{\pi/4}f - 1)$ are the mean size and interparticle distance, respectively, $d$ is the average size of precipitates, $f$ is the volume fraction of precipitates.
Particles in the range of several to thousands of atoms in diameter (i.e., 10-300nm) are effective at impeding dislocation motion. Precipitate strengthened alloys are processed through a two step processing sequence. First, the alloy is heated to a temperature above the solvus and then quickly quenched to promote either no or limited homogeneous nucleation of the precipitate phase. Next the material is heat treated for a set temperature below the solvus temperature to promote nucleation and growth of particles until peak aged condition is achieved. When processing precipitate strengthened materials, it is often observed that first strength increases with precipitate size as volume fraction increases due to the nucleation of new precipitates. This is because larger precipitates ($d$) act as more effective dislocation pinning sites and increases in volume fraction result in decreases in interparticle spacing ($\lambda$). However, when the volume fraction of precipitates reaches equilibrium (peak age), increases in precipitate size result in an increase in interparticle spacing resulting in a decrease in strength. This phenomenon is termed overaging $^{19}$.

**Al-Mg Solution Strengthening.** Al-Mg forms the solid solution in the range 3.5 - 5% Mg at room temperature $^{67}$. Even though the Al-Mg phase diagram (Figure 3.3) indicates limited equilibrium solubility of Mg in Al $^{67}$, the kinetics associated with the formation of the $Mg_{28}Al_{45}$ intermetallic are such that in the compositional range of interest for 5xxx series aluminum alloys the material is easily processed to produce an alloy with Mg in solid solution $^{32}$.

In solid solution, strength is the result of the interaction between the stress field of a glissile dislocation and the stress field generated due to the mismatch between solute and solution atoms sizes/electronegativity/etc. The effectiveness of the solutes as obstacles for dislocation motion can be mathematically simplified to relate Yield Strength ($\sigma$) as
proportional to square root of solute:\n\[ \Delta \sigma = G b \epsilon^{\frac{3}{2}} \sqrt{c} \]
where \( c \) is the concentration of the solute atoms, \( G \) is the shear modulus of the matrix material, \( b \) is the magnitude of the Burger’s vector, and \( \epsilon \) is the lattice strain due to the solute.

### 3.3 Quantification Techniques

#### 3.3.1 Compositional Gradient

Elemental composition is a relative term, which depends on the tool you use to quantify\(^6\). For this study, a comparative study was conducted between three different spectroscopy techniques: quantitative X-ray Energy Dispersive Spectroscopy (XEDS), qualitative XEDS, & X-ray Photoelectron Spectroscopy (XPS). The goal of this comparative study was to assess the suitability of each technique prior to data collection for statistical analysis.

**Quantitative XEDS versus Qualitative XEDS.** XEDS is an electron spectroscopy technique that utilizes characteristic X-rays emitted by the sample when subjected to incident electron radiation. XEDS can be conducted in either a quantitative or a qualitative fashion. In both cases, the position (energy) and relative intensities of the characteristic X-ray peaks are used to quantify the composition of an unknown sample.

Intensity peaks are first identified using reference spectra (which is basically characteristic line shape information). Often, the intensity peaks overlap (e.g., Al K\(\alpha\) - 1.486
Methods and Data Collection

KeV & Br Lα (1.480 KeV) due to low resolution (approx. 140 eV) of XEDS. Therefore, reference spectra help to isolate the correct element by checking for other characteristics peaks from the same element in an unknown spectra. Once identified, then, intensity peaks are compared against the standard spectra of that particular element for quantification. In an ideal case, this comparison provides accurate measurements of the compositional values of an unknown specimen, but depending on the elements present in the sample, the intensity of X-rays emitted are affected (i.e., unintentional fluorescence or absorption events). In 1951, Castaing\textsuperscript{22}, quantified these effects (aka matrix effects), and made it possible to use XEDS for analytical analysis.

Qualitative XEDS uses pure element standards and conduct Castaing\textsuperscript{22} corrections. This gives results, but with errors of approximately 1 at\% or less due to the mismatch between the composition of the standards and composition of the unknown sample (expounded in Castaing\textsuperscript{22}). In quantitative XEDS, this error is reduced to 0.1 at\% or less, by recalibrating the system using standards which have similar composition as that of the unknown sample\textsuperscript{74,75}.

Requirement of standard spectra in quantitative XEDS makes quantification an iterative and long process. The procedure for the selection of an alloy reference sample for quantitative XEDS is as follows: first, the approximate composition of the unknown sample is measured, for example by qualitative XEDS. Next, an alloy sample of known composition must be produced, validated and prepared. The validation of the composition of this alloy sample requires accurate, time consuming techniques such as XPS. Once the alloyed calibration sample is prepared and validated, the XEDS system can be recalibrated and the accuracy of results can be measured. Then, when conducting quantitative XEDS analysis on the unknown sample, reference spectra of the elemental and
known alloy samples must be collected at the same time, under the same conditions, as the spectra of the unknown sample, which puts a time constraint. Below is the Table 3.1 describing the conditions.

<table>
<thead>
<tr>
<th>Spectrum Type</th>
<th>Same Beam Energy</th>
<th>Known probe current</th>
<th>Known live time</th>
<th>Same element / elements</th>
<th>Same detector</th>
<th>Same detector position</th>
<th>Same detector resolution</th>
<th>Same detector calibration</th>
<th>Unobstructed views of the characteristic lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Reference</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3.1. Conditions needed for Quantitative XEDS (Reproduced from Ritchie 74, 75).

In qualitative XEDS, standard spectra on similar equipment are pre measured by the manufacturer of the equipment. The use of these pre-measured spectra produces an error, validated by the manufacturer of 1 at%. For this study, we would require at least two alloyed standards to utilize quantitative XEDS instead of qualitative XEDS. Due to the limited time constraints and funding of the project, it was decided to attempt statistical analysis on the gradient sample using qualitative XEDS.

**Qualitative XEDS verses XPS.** XPS 64,78 use photoelectrons for chemical analysis. With XPS, precision increases by more than 100 fold, but the signal comes from the top 5-10 nm of thickness, making it a surface technique. Having Al as the major constituent, which easily forms an oxide layer on the surface, sputtering became a necessity in case of XPS as opposed to XEDS, where signal comes from top 1 µm depth. In comparison, XEDS is fast, and does not need surface preparation prior to taking measurements, as long as the sample is flat. Lateral resolution of XPS is low, so it is necessary to do large area scans instead of point measurements (XEDS), which increases time and cost,
and also reduces the number of independent observations along the gradient. XEDS lateral resolution can be up to 15nm at 15KeV, which allows very location-specific measurements, which is the key requirement for quantifying compositional gradient. Table 3.2 provides a summary of the experimental constraints for the three techniques explored in this discussion. Based on the table below, it can be seen that qualitative XEDS proves to be the most promising technique for high-throughput experiments.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Excitation Source</th>
<th>Output Signal</th>
<th>Lateral Resolution*</th>
<th>Depth Resolution*</th>
<th>Chemical Analysis Precision</th>
<th>Average Time Span</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative XEDS</td>
<td>Electron Beam</td>
<td>X-rays</td>
<td>15nm at 15KeV</td>
<td>1 µm</td>
<td>1 at%</td>
<td>Few seconds</td>
<td>No Surface Preparation</td>
</tr>
<tr>
<td>XPS</td>
<td>X-rays</td>
<td>Photoelectrons</td>
<td>10 µm</td>
<td>5nm</td>
<td>0.01at%</td>
<td>15-20 minutes</td>
<td>Need Surface Preparation</td>
</tr>
<tr>
<td>Quantitative XEDS</td>
<td>Electron Beam</td>
<td>X-rays</td>
<td>15nm at 15KeV</td>
<td>1 µm</td>
<td>0.1 at%</td>
<td>Few minutes to hours</td>
<td>Need Standards and Reference Spectra</td>
</tr>
</tbody>
</table>

Table 3.2. Comparison of different techniques for chemical analysis. “*” means values are approximate. Actual value might change depending on the material and energy of incident beam.

### 3.3.2 Microhardness

Microhardness measurements provide insight about the strength of the material, specifically a convoluted measure of the yield strength and strain hardening exponent. During indentation, the material deforms plastically under a pre-defined load. A hardness number (i.e., Vickers, or HRC) is a measure of the material resistance to plastic deformation under load based on the shape of the indenter and the size of the indentation.

Suitable load for the microhardness measurements depends on the grain size. If the grain size is larger than the indent size, then the hardness indent is measuring "single crystal" type behavior resulting in large variability in the data from one grain to another.
due to variations in strength and hardening rate as a function of crystallographic orientation\textsuperscript{34,83}. This is the scientific principle behind nanoindentation studies that can measure the modulus and yield strength of individual grains\textsuperscript{53}. If the indent size is significantly larger (3-5 times) than the grain size, then the indent is measuring an average "continuum" property of the sum of all grains and their neighboring interaction. In this case, variability is low in measurements.

**Basic Formulations\textsuperscript{29}.** Developed as an alternative to other macro-hardness measurements such as Brinell and Rockwell, the Vickers test allows for accurate measurements across a larger variety of load scales by employing a pyramidal diamond indenter\textsuperscript{81,82}. Hardness tests can be conducted in either load controlled (most common), or depth controlled mode. The formulations below are derived from the load controlled standpoint, but formulas exist to translate between the two\textsuperscript{29}.

A hardness value is measured by the ratio of force $F$ required to produce an indent with a given surface area of $A$. The Vickers hardness number ($H_v$) has the units of $kg/mm^2$, while DPH (Diamond Pyramidal Hardness) numbers are reported in $N/mm^2$. The area of an indent is given as a function of the diagonal lengths and the included angle of the indenter tip, $\alpha$, which is $136 \pm 0.5^\circ$:

$$
A = \frac{d^2}{2\sin(\alpha/2)} = \frac{d^2}{1.8544}
$$

Therefore, the Vickers hardness, $H_v$ in $kg/mm^2$ is equal to

$$
H_v = \frac{F}{A} = \frac{2Ps\sin(\alpha/2)}{d^2} = \frac{1.8544F}{d^2}
$$
where $P$ is the load in kgf. In base SI units, or the DPH number:

$$H_v(DPH) = \frac{F}{A} = \frac{0.1891F}{d^2}$$

where $F$ is given in Newtons and $d$ in mm. To calculate the depth at which the indenter penetrates the sample surface, the expression for $A$ as a function of depth $h$ is given by:

$$A = 4h^2 \tan^2(\alpha/2)$$

The hardness is a function of this depth $h$ according to:

$$H_v = \frac{F}{24.5h^2}$$

### 3.4 Experimental Specifications

#### 3.4.1 Sample Preparation

Samples (19x19x4 mm) were machined from the plate such that the compositional gradient (short transverse) was parallel to one of the edges (labeled as $z$-direction). Then, they were mechanically polished through 1 $\mu$m diamond paste using Allied Multiprep System, followed by a final vibratory polish using 0.05 $\mu$m colloidal silica.

#### 3.4.2 Framework for Data Collection

Before collecting the data, an implicit framework was placed on the samples to mark locations (Figure 3.4). X-axis along the rolling direction, and Z-axis parallel to the short transverse direction (i.e., the gradient direction). The bottom left corner was marked as the origin, and location measurements were recorded in mm.
3.4.3 Data Collection from Microhardness Measurements

Cursory analysis of the grain size in the gradient regime was conducted utilizing EBSD under the following analysis conditions. Calculation of average grain size assuming a grain boundary has a minimum of $5^\circ$ misorientation and elliptical shape was done using Tango (a software for orientation map display and manipulation). The long axis of the grains, averaged 0.5mm in length, was parallel to the rolling direction and the short axis of the grains, averaged 40 $\mu$m long, was parallel to the short transverse direction (i.e., the gradient direction) at the soft side. This indicates that the material might be a suitable candidate for the measurement of "average properties" via hardness measurements if the size of the indent produced by a defined load is greater than 40 $\mu$m. As such, there is no exact multiple number suggested in ASTM E384-10$^{29}$, a study conducted to choose the suitable load.

A Buehler 1600-4963 indentor was used to measure Vickers hardness on polished samples. To assess the variability in the measurements due to a potential transition between continuum and grain dependent behavior, 22 microhardness measurements (at
each load) were collected in the region with the largest grains at loads ranging from 100 to 1000 gf. The average and standard deviation of the 22 measurements at each load are presented (in Figure 3.5), which highlights that the variation in the 300 gf load is the minimum. To confirm, again 22 measurements were collected one more time at 300 gf load, represented by GF300_2, which also showed the similar variability. Therefore, all subsequent data for model training and testing were collected at 300 gf load.

Figure 3.5. Box Plot of the variation in the hardness measurements at different loads as calculated from 22 measurements at each loading condition.
Microhardness is a destructive technique, which creates an affected zone (or plastic zone\textsuperscript{11}) in space, depending on the applied load and size of the indent, and on properties of the material (e.g., ductility), which hampers the repeatability of experiment in that affected zone. Therefore, when repeating experimental measurements, the affected zones of subsequent indents should not overlap. Empirically it is assumed that for ductile materials, such as aluminum, the affected plastic zone around an indent is 1.5 times greater than the indent size. Therefore, subsequent indents are collected at a minimum distance of 3\times the maximum indent size.

First indent location was marked at (X,Z) \rightarrow (3.175,3.175), and then subsequent indents with 2.54 mm intervals (greater than 165\mu m which is 3\times the indent size at 300gf load) in both directions (e.g., (5.715, 3.175) is the next location along the X-axis, and (3.175, 5.715) along the Z-axis) (Figure 3.6). In total, there were 6 locations along the X-axis, and 6 locations along the Z-axis, resulting in 36 independent observations of performance covering the spatial range (3.175 - 15.875) along both directions. These observations allow for the quantification of measurement variability in regions of "equivalent" microstructure and the the quantification of the variability of performance along the concentration gradient.

### 3.4.4 Data Collection from Qualitative XEDS

The compositional gradient was characterized using qualitative XEDS in an FEI Nova NanoLab 200 at 10KeV energy. Data was collected at the same marked locations, (now, physically marked by indents), and twice at each location (tells about the variability in XEDS) (Figure 3.6). All parameters (like Working Distance, Accelerating Voltage, Current,
Methods and Data Collection

Figure 3.6. Schematic showing how the data is collected for training dataset. Left bottom (soft side or 5xxx surface) was taken as a reference point. Micro-hardness measurements were done at an interval of 2.54mm in both directions, and XEDS measurements at the neighboring positions (twice at each location).

Dwell Time, and Dead Time) in FEI Nova NanoLab 200 are held constant while taking XEDS measurements. These parameters are presented in Appendix A.
4 Exploratory Data Analysis and Semi-gSEM Performance Modeling

Exploratory data analysis (EDA) is an approach to explore the data prior to formal modeling or hypothesis testing. The objectives of EDA\textsuperscript{13,85} are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

4.1 Basic Steps in EDA

4.1.1 Data Collection

The first step in the data collection process is to answer the fundamental questions: what is the goal of the project, and what would be the appropriate data needed to achieve that goal. A clear understanding of the goal and data needed provides context by which
associated variables can be selected. Once data and variables are selected, new questions emerge: Can data be gathered from existing sources / databases, and is it necessary to collect new information through other means. Depending on how the data was collected, limitations of the generalizability of results can also be defined.

The goal of this work was to determine mapping functions and pathway relationships between $<P_r|M|P_e>$ variables. Based on the domain knowledge presented in Chapter 3, variables for microstructure and performance were selected. In this case, the gradients in composition (using qualitative XEDS) and performance (using Vickers hardness) were measured at regular intervals along the short transverse direction. For this study, additional metrics of the microstructure (i.e., grain size and texture) were assumed constant, and no heritage data was incorporated.

### 4.1.2 Data Cleaning and Munging

Data requires tidying before it can be subjected to analysis. Here, it entails assembling and arranging the data into a predefined data structure that allows for approximately coincident observations of multiple variables at equivalent systems. For gradient sample, coincident observations are in space, where multiple observations were taken along the gradient, and equivalent systems were defined by approximate equivalent composition.

For analysis, collected data and metadata were stored into comma-separated values (.csv) files. Metadata (data about the data) included the following: the specific location of each indent, both diagonals length, and other improtant information, for example, date and time of the each measurement. The column structure used to store the data and metadata is presented in Appendix A. Collectively, this process is called as
data munging. While munging the hardness data into csv files, indentations were assumed to be point measurements so they can be stored as point measurements, and not as spatial measurements. Elemental composition data are also considered as point measurements when saved into csv files, and are linked through their positions to their respective hardness measurements.

The electron beam can be focused to 15nm radius on surface of the sample but acquires a larger sub-surface volume. The volume acquired in a single XEDS measurement depends on the energy of the electron beam and the effective atomic number of the material. This is called as interaction volume and can be easily calculated using the Monte Carlo simulation (using software like Casino\textsuperscript{36,47,48}). While data collection, the interaction volume should not overlap from two different measurements, as otherwise, this leads to collecting the same information. Here, XEDS measurements are separated by more than 50 µm (approximate size of indent), which is more than the size of interaction volume for Al at 10KeV (approximately 1 µm (example in Hovington et al.\textsuperscript{47})). Once arranged, EDA is used to search for inconsistencies and to determine a preliminary set of variables that are to be statistically examined for their inclusion as important system components.

4.1.3 EDA & Visualization

A pair-wise plot is presented in Figure 4.1. As previously explained in section 3.4, hardness and compositional variables were measured at 6 different x-locations for each z-location along the short transverse direction. This simple EDA visualization approach provides an opportunity to identify significant variables and visualize potential trends.
between the variables. The trends in the data are both visualized by providing a graphical representation of the variables in the lower anti-diagonal and a mathematical representation of the trend are provided as the r and p-values of a linear correlation function between the variables in the upper anti-diagonal. The r and p-values were calculated with the R-package "stat" for a linear correlation between variable pairs.

In pairs-wise plots (Figure 4.1), the x and y axes of each box in the lower anti-diagonal can be read from the diagonal elements. The variable on the y-axis of each graph in the lower anti-diagonal is defined by the variable listed horizontally on the diagonal, and the variable on the x-axis of each graph is defined by the variable listed vertically on the diagonal. Similarly, r and p-values in the upper anti-diagonal were determined in a mirrored fashion. For example, the graph located in the [2,6] matrix location (assuming [1,1] is the top left) has the horizontal and vertical axis of Z-location and Zn(at%) respectively, and the r and p-values presented in the [6,2] matrix location provide a mathematical representation of a linear fit between these two variables.

From the Mg(at%) - Z(mm) plot, and the Zn(at%) - Z(mm) plot, it's evident that the gradient in composition is continuous along the short traverse direction (Z-axis). Also, while there is a continuous change from Al-5456 (Mg rich) to Al-7055 (Zn rich), the concentrations of Al(at%) and Cu(at%) do not change, and therefore would not likely provide significant components for analysis. Hardness with Z(mm), shows a continuous increase as the gradient change from soft Al-5456 to hard Al-7055. Coupling hardness with Mg(at%) or Zn(at%) (matrix location [5,3] and [6,3]), shows high variation, reflecting the aggregate sum of noise levels of two different techniques used.
4.1.4 Interpreting Results

At this step of the data-science approach, the hypothesis can be altered based on the observed trends, and appropriate statistical tools and techniques are selected to determine
the statistical significance for the assumed hypothesis. For this study, for example, the domain knowledge led to the initial hypothesis that the gradient in composition governs the gradient in performance. The EDA observed trends may suggest an alternative or modified hypotheses. From Figure 4.1, it’s evident that Hardness is increasing inversely with Mg(at%) and proportionally with Zn(at%) composition. This validates the initial hypothesis based on the domain knowledge and provides the opportunity to assess the validity of inverse and proportional functional forms.

4.2 Semi-gSEM

A new domain-guided network modeling approach was employed for this study. This approach provides a statistical framework to understand, predict, and design the performance of materials and material systems from experimental measures of metrics of performance and microstructure evolution.

Step-wise regression was used to move through the network, from microstructure and performance variables in a Markovian spirit, whereby the relationship amongst two variables in the regression is unaffected by prior analysis, and other variables are considered to be independent constants. The pathway (edge) relationships among the nodes (vertices) in the network are evaluated against a set of permitted functional forms (Table 4.1). Then the relationships between variable pairs are rank-ordered to determine their significance to assess the best functional forms for each variable pair via the statistical measures of \( \text{adj-} R^2 \). The final step is the report of the system of equations for the pair-wise variable relationships and the full semi-gSEM network, so that the semi-gSEM model can be visualized on the dataset for training, testing and validation.
### 4.2.1 Role of Functional Forms

Functional forms can be added or removed, based on the domain knowledge. For example, in solute strengthening, strength is correlated with the concentration of the alloying additions by the following governing equation \(^{31}\):

\[
\Delta \sigma = G b c^{\frac{3}{2}} \sqrt{c}
\]

where \(c\) is the concentration of the solute atoms (in this case Mg concentration), \(G\) is the shear modulus, \(b\) is the magnitude of the Burger’s vector, and \(\epsilon\) is the lattice strain due to the solute. If it is assumed that the \(\epsilon\) term is constant for a particular solute atom type; and that \(G\) and \(b\) are also constant; then, this direct correlation gives us a functional form \(\text{SQR}\) (Table 4.1) for semi-gSEM quantification.

For precipitation hardening (or precipitate strengthening), it is difficult to simplify the complexity of the model (presented in section 3.2.1) without any insight into the
complex relationship between processing parameters (unknown) and the variable measured in this study (composition), as composition and processing parameters will control volume fraction and precipitate size. Therefore, an unbiased assessment with semi-gSEM is used to assess the correlation of Zn(at%) composition with strength by this mechanism.

4.2.2 Semi-gSEM Quantification

Semi-gSEM step-wise regression package (v0.4.5)\textsuperscript{17,18} was used to rank-order statistical significance between each variable-pair for each of the functional forms given in Table 4.1, by comparing $\text{adj-}R^2$ values. Semi-gSEM quantification indicated that an exponential relationship between Zn concentration and Hardness, and an inverse square relationship between Mg concentration and Hardness had the highest statistical significance. The latter functional fit was contradictory to the domain knowledge for solute strengthening, where it was expected that hardness would be related to the solute concentration thru a square-root relationship. This mismatch could be the result of high variability in datasets. Therefore, a confirmatory analysis was conducted, which manually assessed the $\text{adj-}R^2$ values for different functional forms. This manual assessment is presented in Figure 4.2, and it highlights the similarities between the top three ranked functional forms: $\text{adj-}R^2$ for Log: 0.5707, $\text{adj-}R^2$ for Square Root: 0.5703, & $\text{adj-}R^2$ for Inverse Square Root: 0.5707. A simplified network model derived from the biased SEM quantification is shown in Figure 4.3, with Square Root functional form for relationship between Mg and strength.
4.3 Model - Training/Testing/Validation

After quantifying the observed trends with the semi-gSEM regression package, a domain-supervised, biased, statistical analysis was conducted to test the statistical validity of the assumption that the two independent strengthening mechanisms contribute in concert: 

\[ H = \alpha e^{Zn} + \beta \sqrt{Mg} + \gamma \]. For this analysis, the coupling relationship with the highest significance between Hardness and Zn concentration was selected because the domain knowledge does not allow to simplify the equations to a single variable relationship. The diagnostic model provides a representation of the performance variable with regards to the compositional variables with a statistical significance of an adj-\( R^2 \) of 0.6, and coupling coefficients as follows: \( H=0.24e^{Zn} - 29\sqrt{Mg} + 247 \).
High variability in Figure 4.4, confirms low adj-\(R^2\) value of predictive model (for training dataset - 0.60). To check, whether the assumed functional form \(H = (0.24e^{Zn} - 29\sqrt{Mg} + 247)\) is symmetric, or is biased, residuals were plotted (Figure 4.5).

Coupling coefficients were calculated using least-squares regression which assumes 1) that the residuals are normally distributed, and 2) that residuals are centered such that their variance doesn’t change as a function of the dependent variable\(^{86}\), therefore, it is beneficial to check residuals. There are four methods of visualizing the residuals to check if they are consistent with the assumptions for least-squares regression. For definitions of statistical terms, please check Appendix B.

The first method of visualizing the residuals is to plot the residuals verses the fitted values (or predicted values). This gives an idea of whether there is any systematic variation of the residuals with the fitted values. If the red line has a slope of zero, with an intercept of zero, then the residuals are symmetrically distributed. If the red line is
strongly curved, or if there is any apparent pattern in the points on this plot, then the model may not be the appropriate one. In this case, the curvature is not strong. Points that tend towards being outliers are labeled (19,20,71 in this case).

The second plot is the normal Q-Q plot, which plots theoretical quantiles of fitted values against standardized residuals. This is to check whether the residuals are normally distributed. If the residuals were normally distributed, this visualization would produce a straight line with a slope of one. Some deviation is to be expected, particularly near the ends (note the upper right), but the deviations should be small, as they are here.

The third plot is used to check if the variance is constant (i.e., if the standard deviation among the residuals appears to be about constant). If the red line is strongly tilted
Figure 4.5. Residual Plots of functional form - $H = 0.24e^{zn} - 29\sqrt{Mg} + 247$.

up/down, that it is a red flag. There are no issues - the variance appears constant. (The red line will always move up/down a little because of inherent randomness).

The last plot (residuals vs leverage) is used to check if there were any overly influential points. Leverage of an observation is based on how much the observation's value
differ from the mean of the predictor variable, which tells how much a particular observation influences the regression. Standardized residuals from a normal distribution should be centered around zero and range between 2-3 standard deviations away and symmetrically. Because the regression must pass through the centroid, points that lie far from the centroid have greater leverage, and their leverage increases if there are fewer points nearby. As a result, leverage reflects both the distance from the centroid and the isolation of a point.

This plot also has contours values of Cooks distance, which measures how much the regression would change if a point was deleted. Cooks distance is increased by leverage and by large residuals: a point far from the centroid with a large residual can severely distort the regression. On this plot, you want to see that the red smoothed line stays close to the horizontal gray dashed line and that no points have a large Cooks distance (i.e, dashed line at 0.5). Note, one isolated point towards the far right has high leverage, but it’s residual is close to zero indicating that it would have limited affect on the regression if were removed from the analysis. From these plots, it’s fair to say that the functional form is symmetric.

To validate this diagnostic model, same datasets were collected from a different sample with different data density (3 Z-levels with 11 different locations at each Z-level, which changed the data intervals from 2.54 mm in both directions in training datasets to 1.27 mm and 6.35 mm in the X-direction and Z-direction, respectively (Figure 4.6)). This is called as “test dataset”. Using the model equation, predicted values of Hardness measurements for the test dataset were recorded and plotted. In Figure 4.7, (top) actual and predicted values were overlaid on Z-axis which shows that the variation in the actual values and predicted values at a particular ‘Z’ is approximately same. (Bottom)
Figure 4.6. Pairs-wise plot and linear correlation coefficient between atomic percentage of (Al-Mg-Zn-Cu) in the composite and Hardness for the test data set. X and Z represents the same local co-ordinate system as defined for Figure 4.1. r and p values in the upper right quadrant are linear correlation coefficients.

By normalizing the actual values along X-axis, scatter in predicted values were shown, which is more or less symmetric. High variance in predicted values is inherent from the training and testing datasets. From this, it can be inferred that the model is predictive within the range of study.
Figure 4.7. Predicted hardness values (red) based on measured compositional gradient in comparison to measured hardness values. (Top) Overlapped with z-axis. (Bottom) Scatter in predicted values.
5 Discussion

Data-driven semi-gSEM network modeling provides quantified equations of the relationships between variables, but it is essential to consider limitations of this new approach, since they can inform its appropriate application. Figure 4.4 does indicate that the functional forms provide a diagnostic understanding of the mechanisms of strengthening but the trends observed are counter to the simple assumption that increasing solute concentration increases hardness. This arises from the fact that even though strengthening mechanisms are independent, the changes in concentration are not independent. The EDA in Figure 4.1 indicates that increase of Mg concentration is at the expense of the Zn concentration. Therefore, one needs to be cautious when extrapolating the network model to different concentrations of Mg and Zn.

For example, Al-5059 alloy contains 6-6.5 at% Mg and 0.2-0.3 at% Zn (rest Aluminum). For this alloy, the measured hardness values was in the range of 114-117.7 VHN at 300gf while the network model predicts a hardness of 173.5-175.5 VHN.

Let’s look at the diagnostic model: \( H = (\alpha e^{Zn} + \beta \sqrt{Mg} + \gamma) \), where \( \alpha \) is 0.24, \( \beta \) is -29, and \( \gamma \) is 247. Mathematically, here, functional forms (for e.g., Exp & SQR) captures the curvature and coefficients \( \alpha \) and \( \beta \) captures the nature of relationship (for e.g., direct or inverse), however, \( \gamma \) (or intercept) is meaningless. Let’s assume a case, where Zn(at%)
and Mg(at%) are both zero, which means that the sample is pure aluminum with little copper, with a hardness value of 247, which is inconsistent with domain knowledge, as it says that the materials strength is significantly increased when we take out the Mg and Zn. This is the same reason for a relatively high predicted value for Al-5059 alloy. The purpose of linear regression in scientific research is just to understand the relationship between predictor (here Mg and Zn) and response (hardness).

This diagnostic model, suitably captures 60% of the variance (or adj-$R^2$ is 0.6) with two variables. Rest, 40% of the irreducible error is attributed to multiple factors. One of the factor is uncertainty in experimental measurements. Buehler 1600-4963, used for microhardness measurements, involves human interaction for measuring the diagonals length of indent, which invokes a mismatch in every measurement due to limited resolution and fixed magnification of the magnifying screen. The inherent nature of the model equation is such that it assumes all other factors as constant, which isn't the case. From preliminary EBSD analysis (done for calculation of grain size), it was implied that microstructural properties change with compositional change, and these structural properties also affect the strength. Incorporating these latent variables might reduce the irreducible error from the regression.

Further, “how the data was collected to train and test the model”, also give insights about the things which can make the model more generic. Like, in the case of microhardness measurements, having a more versatile way of measuring strength, which isn’t dependent on scale (like VHN or HRC) and load (from 100 gf to 1000 gf), where things are comparable across wide range of materials would be better (e.g., tensile strength). In case of XEDS results, changing the XEDS spectroscopy parameters change the results, and they need to be altered as the material is changed for better precision. For example,
if some heavy element is present in the alloy (like V), beam energy should be sufficient
to knock off characteristic X-rays for these heavy elements.
6 Conclusions

In conclusion, we have derived a robust framework for integrating different experimental metrics for quantifying local variables of composition and performance. We show both unbiased and biased statistical analysis approaches to assess the strength of coupling coefficients in functional forms so as to develop a diagnostic model that provides both predictive capabilities and validates existing theories. The domain-guided diagnostic model was trained and validated over a limited compositional range, showing statistical significance with an adj-$R^2$ of 0.6. The low adj-$R^2$ value is attributed to the large degree of scatter in the hardness variable.

Future work can assess if the degree of scatter can be attributed to other microstructural variables such as grain orientation, grain size, aspect ratio, Taylor factor. Additionally, semi-gSEM approach needs to be modified to accommodate functional forms that can account for the interaction between variables before applying the technique to more complicated <Processing|Microstructure|Performance> relationships.
Appendix A

Databook

Description of variables, metadata and bitbucket repository.

1 For Hardness Measurements

- **Metadata**
  - DateTime: Date & Time of the measurements
  - SpecimenID: A generic sample ID
  - HardnessScale: represents the Hardness Scale, VHN is Vickers Hardness Number
  - LoadApplied: Load Applied for indentation in gram force (gf)
  - Diagonal.1: Diagonal Length of Indent, in micrometers
  - Diagonal.2: Second Diagonal Length of Indent, in micrometers

- **Variables**
  - disp.ref.z: location of Origin on z-axis with respect to reference co-ordinate system, in mm
  - disp.ref.x: location of Origin on x-axis with respect to reference co-ordinate system, in mm
  - dz: step size on z-axis, in mm
  - dx: step size on x-axis, in mm
  - Location.Z: local co-ordinate z-axis, each value is a whole number
  - Location.X: local co-ordinate x-axis, each value is a whole number
– Location.Y : local co-ordinate y-axis, each value is a whole number, and single value represents a single sample
– Hardness : Hardness Value in VHN

2 For XEDS Measurements

• Metadata
  – DateTime : Date & Time of the measurements
  – SpecimenID : A generic sample ID
  – BeamCurrent : Beam Current in nano Amperes (nA)
  – AccVoltage : Accelerating Voltage in Kilo Volts (kV)
  – WD : Working Distance, in mm
  – AcquireTime : Time taken to acquire spectra, in seconds
  – DeadTime : Dead time during the collection of spectra, in % of the total time
  – Magnification : Magnification of microscope at the time of spectra
  – Num.Channels : Number of Channels in the spectra

• Variables
  – disp.ref.z: location of Origin on z-axis with respect to reference co-ordinate system, in mm
  – disp.ref.x: location of Origin on x-axis with respect to reference co-ordinate system, in mm
  – dz : step size on z-axis, in mm
  – dx : step size on x-axis, in mm
– Location.Z: local co-ordinate z-axis, each value is a whole number
– Location.X: local co-ordinate x-axis, each value is a whole number
– Location.Y: local co-ordinate y-axis, each value is a whole number, and single value represents a single sample
– Al: Amount of Aluminum, in weight numbers (wt), at the location - later normalized to 100%
– Cu: Amount of Copper, in weight numbers (wt), at the location - later normalized to 100%
– Mg: Amount of Magnesium, in weight numbers (wt), at the location - later normalized to 100%
– Zn: Amount of Zinc, in weight numbers (wt), at the location - later normalized to 100%

3 Statistical Framework

A framework was designed to navigate all the data collected for different variables. For combining different datafiles, collected at different time and at different instruments, some common attribute was needed to link all of them.

As all of the data was location driven, e.g. composition was driven by spatial location because of the designed gradient, an implicit framework was laid to mark all the locations on the sample. As every measurement acquires certain spatial volume, depending on the other attributes, e.g. load in hardness measurements change the indent size, it's not possible to have measurements at every observed location. Proper intervals were decided, so no two measurements affect each other.
To capture different data strength (e.g., number of observations per square millimeters), a method was employed containing a local coordinate system and a global coordinate system (aka reference coordinate system). Local coordinate system to capture the different data strengths among different datasets, and a reference coordinate system for combining all datasets into a single large dataset, which enables navigation for EDA.

Each local coordinate system defined by its origin (disp.ref.x, disp.ref.z) and step size (dx, dz). For reference co-ordinate system, a fixed origin (left corner at the soft end) was defined, which was same across different samples, and all measurements in real length scales (in millimeters). Reference coordinate system had an advantage that every sample had the same dimensions. Therefore, irrespective of the data strengths in different files, once the data is in reference coordinate system, different samples or different datasets among the same sample were comparable using location as the marker.

Different datasets were collected in different csv files. These csv files had all the attributes, listed in above section, in different columns, and different measurements in different rows. Once all the data were collected, it was ingested into R (an open source, statistical computing platform), where every dataset transformed to reference coordinate system. While ingestion of data, some assumptions were made, for example, as discussed earlier in chapter 3, microhardness indents and XEDS measurements were treated as point measurements, considering that no two measurements were affecting one another.
4 “1602-ARO-Data” Repository

“1602-ARO-Data” is a bitbucket (https://bitbucket.org) repository containing all the data, figures, and codes, used for the analysis. Repo link - https://bitbucket.org/aak126/1602-aro-data.

Repository contains a folder structure for sharing different aspects of the study. Folder “data” contains all the data files, “figs” contains all the figures generated through R, and “Rmd” contains all the Rmarkdown files. Rmarkdown files are self explanatory as they include detailed comments for every respective code chunk. “.R” files are R script files and they contain only codes.
Appendix B

Statistical Concepts

Definitions and mathematical formulations of concepts used in building diagnostic model.

1 Linear Regression

Linear regression does a linear fit between the dependent (y) and independent (x) variable.

\[ y = \alpha + \beta \times x + \epsilon \]

where \( \alpha \) and \( \beta \) are the coefficients and \( \epsilon \) is the error term.

\( \alpha \) and \( \beta \) are evaluated by minimizing the sum of squares of the error term.

2 Residuals

The difference between the observed value of the dependent variable and the predicted value is called the residual. Each data point has one residual. Notice, residual is defined after model is evaluated by minimizing the sum of squares of the error term.

The error (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true value of a quantity of interest (for example, a population mean), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean).
3 Standardized Residuals

The standardized residual is the residual divided by its standard deviation. It’s a measure of the strength of the difference between observed and expected values. With standardized residual makes it easy to see which observations are contributing the most to the value, and which are contributing the least.

4 $R^2$ and adj-$R^2$

$R^2$ is the quotient of the variances of the fitted values and the observed values of the dependent variable, where variance is a numerical measure of how the data values is dispersed around the mean. This gives an estimate of how much variability in the dependent variable is captured by the model.

Adjusted $R^2$ adjusts the statistic based on the number of independent variables in the model. Mathematically,

$$adj\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where: $R^2$ is sample R-square, $p$ is number of predictors, and $N$ is total sample size.

5 Correlation Coefficient

The correlation coefficient of two variables in a data sample is their covariance divided by the product of their individual standard deviations, where covariance is a measure of how much two random variables vary together. Covariance is similar to variance, but where variance tells you how a single variable varies, covariance tells you how two
variables vary together. Correlation coefficient is a normalized measurement of how the two variables are linearly related.

Mathematically, covariance is

$$\text{Cov}(X, Y) = \frac{\sum (X - \mu)(Y - \nu)}{n - 1}$$

where: X is a random variable, E(X) = \mu is the expected value (the mean) of the random variable X, E(Y) = \nu is the expected value (the mean) of the random variable Y, n = the number of items in the data set

Correlation coefficient

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}$$

where: sd(X or Y) is standard deviation of X or Y, respectively.

6 Null Hypothesis and p-value

Null Hypothesis refers to a default position that there is no relationship between two observed phenomena, until proven otherwise. p-value is one way to determine whether a hypothesis is true or false.

The p-value quantifies the strength of the evidence against the null hypothesis and in favor of the alternative. For estimate of p-value, we need certain information beforehand: 1) experiment’s expected results, 2) experiment’s observed results, and 3) degrees of freedom (measure of the amount of variability involved in the research). Then, chi square, numerical value that measures the difference between experiment’s expected
and observed values, is calculated. Then, using chi square and degree of freedom, p-value is estimated using reference tables. One simple example is presented in Figure B.1. These reference table can be further extended for large degrees of freedom.

Once p-value are estimated, they are compared with a pre-defined threshold (called as significance level) for rejecting or accepting the null hypothesis. Traditionally, 0.05 or 5%, is a usual choice for significance level.

<table>
<thead>
<tr>
<th>Degrees of Freedom (df)</th>
<th>Probability (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>0.004</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>0.71</td>
</tr>
<tr>
<td>5</td>
<td>1.14</td>
</tr>
<tr>
<td>6</td>
<td>1.63</td>
</tr>
<tr>
<td>7</td>
<td>2.17</td>
</tr>
<tr>
<td>8</td>
<td>2.73</td>
</tr>
<tr>
<td>9</td>
<td>3.32</td>
</tr>
<tr>
<td>10</td>
<td>3.94</td>
</tr>
</tbody>
</table>

Figure B.1. Chi-square Distribution Table.

7 Quantiles

Quantiles are cutpoints dividing a set of observations into equal sized groups. There are one fewer quantiles than the number of groups created. Thus quartiles are the 3 cut points that will divide a dataset into four equal-size groups.
8 Leverage and Cooks Distance

Leverage is a measure of how far away the independent variable values of an observation are from those of the other observations. The more a data point differs from the mean of the other observations, the more leverage it has. Cooks distance measures the effect of deleting a given observation. Mathematically, cooks distance is

\[
D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{\rho \text{MSE}}
\]

where: \(\hat{y}_j\) is the \(j_{th}\) fitted response value, \(\hat{y}_{j(i)}\) is the \(j_{th}\) fitted response value, where the fit does not include observation \(i\). MSE is the mean square error. \(\rho\) is the number of coefficients in the regression model.
Appendix C

Semi-gSEM Regression Package

Semi-gSEM step-wise regression package (v0.4.5\textsuperscript{17,18}) was used at first for comparison and ranking of functional forms. These functional forms (Table C.1) were designed to study the degradation cycle of acrylic polymers\textsuperscript{17}.

<table>
<thead>
<tr>
<th>Variable Fits</th>
<th>Functional Forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear (L)</td>
<td>$y = a + bx + \epsilon$</td>
</tr>
<tr>
<td>Quadratic (Q)</td>
<td>$y = a + bx + c \times x^2 + \epsilon$</td>
</tr>
<tr>
<td>Simple Quadratic (sQ)</td>
<td>$y = a + c \times x^2 + \epsilon$</td>
</tr>
<tr>
<td>Exponential (Exp)</td>
<td>$y = a + d \times \text{exp}^x + \epsilon$</td>
</tr>
<tr>
<td>Logarithmic (Log)</td>
<td>$y = a + f \times (\log(x)) + \epsilon$</td>
</tr>
<tr>
<td>Linear Change Point</td>
<td>$y = a + b \times x + b_1 \times (x - c) + \epsilon$</td>
</tr>
<tr>
<td>Nonlinearizable Exponential (G-up, H-down)</td>
<td>$y = a + d \times (1 \pm \text{exp}(g(x - h))) + \epsilon$</td>
</tr>
</tbody>
</table>

Table C.1. Uni-variate functionals forms evaluated by semi-gSEM (v0.4.5)

For application in metals, a local copy was forked from semi-gSEM repo and functional forms were modified. Repo Link - https://bitbucket.org/cwrusdle/sgsem. This local copy was called as “strength”. Two new functional forms were added, 1) square root - to capture the solute strengthening, and 2) inverse square root - to capture hall-petch relationship (Table C.2)

Semi-gSEM regression builds an object of class sgSEM, which is a list of the following items:

- “table”: A matrix. For each row, first column is the response variable, second column is the predictor, the other columns show corresponding summary information: Best functional form, R-squared, adj-R-squared, P-value1, P-value2
Variable Fits | Functional Forms
---|---
Simple Linear (L) | $y = a + bx + \epsilon$
Quadratic (Q) | $y = a + bx + c \cdot x^2 + \epsilon$
Simple Quadratic (sQ) | $y = a + c \cdot x^2 + \epsilon$
Exponential (Exp) | $y = a + d \cdot \exp(x) + \epsilon$
Logarithmic (Log) | $y = a + f \cdot (\log(x)) + \epsilon$
Square Root (SQR) | $y = a + b \cdot \sqrt{x} + \epsilon$
Inverse Square Root (ISQR) | $y = a + b / \sqrt{x} + \epsilon$

Table C.2. Uni-variate functionals forms evaluated by semi-gSEM strength branch

and P-value3. The P-values correspond to those of estimators of linear regression coefficients. (Figure C.1)

- “bestModels”: A matrix. First dimension indicates predictors. The second dimension indicates response variables. The i-j-th cell of the matrix stores the name of the best functional form corresponding to the j-th response variable regressed on the i-th predictor.

- “allModels”: A three dimensional list. The first dimension indicates predictors. The second dimension indicates response variables. Third dimension indicates the fitting results of all 6 functional forms. The i-j-k-th cell of the list stores a “lm” object, corresponding to the j-th response, i-th predictor and the k-th functional form.

Object of class sgSEM gives other basic information too, e.g., summary (Figure C.2) and full semi-gSEM plots (Figure C.3).

Semi-gSEM plots are a better way to visualize, than the tables (Figure C.1) or summary (Figure C.2).

For more information on semi-gSEM package, please check sgSEM repository.
Figure C.1. Table generated of object of class sgSEM, before and after the addition of new functional forms.
Figure C.2. Summary of the object of class sgSEM, before and after the addition of new functional forms.
Figure C.3. Semi-gSEM network model plots generated from object of class sgSEM using DiagrammeR package.
Appendix D

Preparation of this document

This document was prepared using pdfL\LaTeX and other open source tools. The (free) programs implemented are as follows:

- \LaTeX implementation:
  - MiKTeX
    - http://www.miktex.org/

- For Version Control
  - BitBucket
    - https://www.bitbucket.org/

- Bibliographical:
  - Bib\TeX
    - http://www.bibtex.org/
  - Zotero
    - https://www.zotero.org/
References

[1] The key word in Data Science is not Data, it is Science | Simply Statistics. URL http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/.


Bibliography


[84] Core Team. HKL Channel 5. URL http://www.oxford-instruments.com/.

