



Classification of T6 Tempered 6XXX Series Aluminum Alloys Based on Machine Learning Principles

TANU TIWARI ^{1,2} SADEGH JALALIAN,^{1,3} CHAMINI MENDIS,^{1,4}
and DMITRY ESKIN^{1,5}

1.—BCAST, Brunel University London, Uxbridge, Middlesex UB8 3PH, UK. 2.—e-mail: Tanu.Tiwari@brunel.ac.uk. 3.—e-mail: Sadegh.Jalalian@brunel.ac.uk. 4.—e-mail: Chamini.Mendis@brunel.ac.uk. 5.—e-mail: Dmitry.Eskin@brunel.ac.uk

Aluminum alloys are widely used in each sector of engineering because of their lower density coupled with higher strength compared to many existing alloys of other metals. Due to these unique characteristics, there is acceleration in demand and discovery of new aluminum alloys with targeted properties and compositions. Traditional methods of designing new materials with desired properties, like ‘domain specialists and trial-and-error’ approaches, are laborious and costly. These techniques also lead to the expansion of alloy search area. Also, high demand for recycling of aluminum alloys requires fewer alloy groups. We suggest a machine learning design system to reduce the number of grades in the 6XXX series of aluminum alloys by collecting the features involving chemical composition and tensile properties at T6 tempering state. This work demonstrates the efficiency of grouping the aluminum alloys into a number of clusters by a combined PCA and K-means algorithm. To understand the physics inside the clusters we used an explainable artificial intelligence algorithm and connected the findings with sound metallurgical reasoning. Through machine learning we will narrow down the search space of 6XXX series aluminum alloys to few groups. This work offers a useful method for reducing compositional space of aluminum alloys.

INTRODUCTION

Aluminum (Al) and its alloys provide the unique combination of properties, which makes them economical, versatile and attractive metallic materials for many uses—from highly ductile, soft wrapping foil to the most demanding structural applications.¹

As pure aluminum is rather soft, the alloying elements are used to improve and control the properties in Al alloys. Most common additions are manganese, copper, silicon, zinc and magnesium. Up to 2 wt.% total amount of these elements can be typically present in an Al alloy, with some specialized alloys containing even large amounts of additives. Also, some minor alloying elements are added in the amounts < 0.5%. These elements have a function of controlling some specific properties, e.g.,

recrystallisation or corrosion resistance. The Al alloys are classified as wrought, casting and rapidly solidified/ powder alloys, which are further subdivided as age- and work-hardenable alloys.¹ These classes are further subdivided into various systems based on the selection of alloying elements.

Aluminum alloy design, since the beginning of the twentieth century, has been essentially an iterative and empirical process, based on the lessons learned from experience and in-service use.² A hill-climbing approach is taken in the traditional development and design of Al alloys.³ This traditional method of research and development is laborious, expensive and does not consider the full spectrum of potential properties. Testing of billions of combinations of alloys is not possible.^{4,5} Alloy development by mixing a combination of alloying elements and characterizing their structure and testing their properties is slow, costly and does not fully harness the data that have been accumulated before. Examples of empirical Al-alloy discovery include the

(Received May 4, 2023; accepted July 18, 2023;
published online August 22, 2023)

gradual and continuous development of Al-Cu alloys to Al-Cu-Mg, then to Al-Cu-Zn-Mg and Al-Cu-(Mg)-Li alloys, based on the requirements of structural applications.⁶

Currently, 7 million tons of recycled aluminum scrap and around 22 million tons of new Al are being used. Recycling aluminum has both financial and environmental benefits. The production of 1 ton of primary Al requires 14,000 kWh energy to create.⁷⁻⁹ In contrast, only 5% of this is required to remelt and recycle 1 ton of Al. Ideally, recycled and virgin Al alloys would be of equal quality.

However, the recycling of complex in composition wrought alloys faces a few challenges. Wrought Al alloy production from scrap has the main difficulty of attaining the desired chemical composition with a less addition of pure alloying elements and also of primary aluminum.⁵ Technically, mixing of various scrap leads to the uncontrollable concentration of critical elements such as iron, copper, manganese, magnesium, zinc and silicon.¹² Once the concentration of these critical elements falls outside of the concentration limit for a particular wrought alloy, the only solution would be their dilution with primary Al or modification of alloy composition. Another technical solution is to carefully sort and combine scrap at the pre-melting stage to assure the suitable chemical composition of the batch is costly.^{7,10} Therefore, due to the large search space of alloying elements present and great variety of existing Al alloys, grouping of alloys in the classes would be a potential solution.

Another well-known issue is the “compositional-tolerance limit” existing especially for wrought alloys, which becomes critical in recycling where unusual and unexpected impurities can inadvertently creep in, and even normal impurities may tend to build up and accumulate to an unacceptable degree.¹⁰ In most cases, it is not well investigated but rather the influence of these tolerance limits on the properties (and particularly on the selected properties) is known from experience with wrought alloys.^{7,11} These recycling issues show an urgent need for optimization of Al alloy grades to narrow down their search space.

The optimization of existing alloys and discovery of new alloys until now have been based on thermodynamic modeling¹¹ and ‘ab initio’ methods.¹² However, these methods have limitations in linking the composition to the properties and to the real processing and service conditions. Traditional methods are not suitable for finding new optimal alloys compositions from the large compositional search space.

However, there are several advanced methods like machine learning (ML) and artificial intelligence (AI) that aim to develop self-learning models that can solve these problems, acting like the human brain. Machine learning aims at building statistical models for data analysis and prediction. The artificial intelligence algorithms construct an

inference model connecting the targeted property to material descriptors by successfully learning from the past data and understanding the pattern inside the data, which results in the rational choice of next experiment and makes accurate predictions.¹³

Today, materials science successfully uses machine learning algorithms to solve various materials science problems.¹⁴ The prediction of new stable materials, calculation of multiple material properties and acceleration of first-principle calculations are only a few examples of the numerous machine learning applications in materials research that have already been shown to be effective.¹⁵

For instance, Raccuglia et al.¹⁶ successfully designed new materials by employing a machine learning strategy to learn the rules of material synthesis from failed experiments. High-performance copper alloys with inverse compositional design were accomplished by building a method for designing alloy compositions with a focus on alloy properties; this method was also used to design the composition of piezoelectric materials.¹⁷ By integrating an adaptive learning technique and a machine learning model, Lookman et al.^{18,19} were able to successfully manufacture high-property shape memory alloys. Similar machine learning methods were applied in the compositional design of inorganic superconducting material,²⁰ piezo-electric materials,^{21,22} high-entropy alloys²³ and stainless steel²⁴ as well as in the property and structure predictions such as diffusion,^{25,26} lattice misfit,²⁷ density,²⁸ fatigue,²⁹⁻³¹ Seebeck coefficient,³² glass-forming ability,^{33,34} atomic force field,³⁵ strength,³⁶ elastic constant,³⁷ etc. For the design and prediction of structure and properties, most of the current machine learning algorithms (ML) used the concentration of specific elements as the input/output to construct regression models of composition-property design. Most of these works focused on designing new alloys, which does not help us in narrowing down the composition space of alloys.

Also, there have been many attempts to apply clustering algorithms in materials science problems, e.g., Nenchev et al.³⁸ combined supervised and unsupervised machine learning for predicting and estimating the hardenability in gear steel samples. Golowanov et al.³⁹ devised a recursive algorithm for forming homogeneous groups (clusters) of semiconductor devices based on the maximization of silhouette score by using the K-means clustering algorithm. Kazakovtsev et al.⁴⁰ offered a number of clustering models and techniques to address the issue of dividing a presumably heterogeneous batch of semiconductor devices into homogeneous groups (clusters). Previously, there were many attempts to use clustering but their focus was also not on narrowing down the composition space of alloys.

In the present work, we proposed a design loop for the rational design of alloy clusters based on key element features and tensile properties, which is identified by machine learning and materials

science knowledge. Due to good formability, machinability, weldability and high resistance to corrosion, we have selected 6XXX series alloys as the case study. This series of heat-treatable alloys generally acquires the top strength in T6 temper, i.e., being strengthened by precipitation hardening.¹ In this case study, we narrowed down the solution space by classifying the 6XXX series dataset in groups using a combined approach of clustering and principle component analysis (PCA) as the main algorithms. Also, LIME (local interpretable model-agnostic explanations) algorithm was applied to explain the clusters provided by the combined approach of PCA and K-means algorithms.

METHODOLOGY

To reach the objectives of this study, we propose the following design loop: data collection → combined K-means clustering and PCA analysis → box plot → LIME algorithm → metallurgical reasoning behind clustering. These steps and their objectives are shown in Fig. 1. Note that we limit ourselves to the extruded alloys in T6 conditions and to the tensile properties at room temperature. This approach makes the database manageable for the case study.

Data Collection

Information gathering and dataset training is the initial step of machine learning (ML) workflow, which heavily depends on the objective of the model that we want to train. The data for the 6XXX series Al alloys were collected from literature sources, journal papers and open web searches and saved in Excel file in csv format. For our case, the data included alloy composition and their tensile properties at T6 tempering conditions.¹ For preparing the data for machine learning algorithms, cleaning of

data was performed. Mostly, the alloys with the full set of the required data were chosen. Some approximating terms can be used in machine learning algorithms such as *int*, *float* and *char* corresponding to integer, decimal and character variables, respectively. Also, we used Jupyter Notebook, an application of Anaconda Navigator, to implement all the machine learning algorithms in this work, and the python libraries used for visualization of figures are Matplotlib.pyplot and Seaborn.^{13,41}

Combined Approach of Clustering and PCA

Clustering is an unsupervised machine learning algorithm.⁴² In the k-means algorithm every data point is repeatedly assigned to the cluster with the closest centroid, and then based on the mean of the data points assigned to each cluster, the centroids or cluster centers are updated. K-means clustering algorithm, for several reasons, often does not work well for high dimensionality datasets; hence, to improve the efficiency, prior to clustering we applied principle component analysis (PCA) on the original data set for better visualization of clusters.⁴³

PCA is a dimensionality reduction technique which reduces the dimensionality of dataset consisting of many variables. It works by identifying the directions of highest variance, i.e., principle components in the dataset, and then creating a low-dimensional space consisting of these principle components. The main objective of combining k-means clustering with PCA is to apply k-means clustering to the reduce dimensionality space to identify patterns or clusters. The clusters obtained by this combined method provided the insight into how the data is organized and how different data points related to the others.

There are a few steps involved in this combined method to better visualize our 6XXX data set:⁴⁴

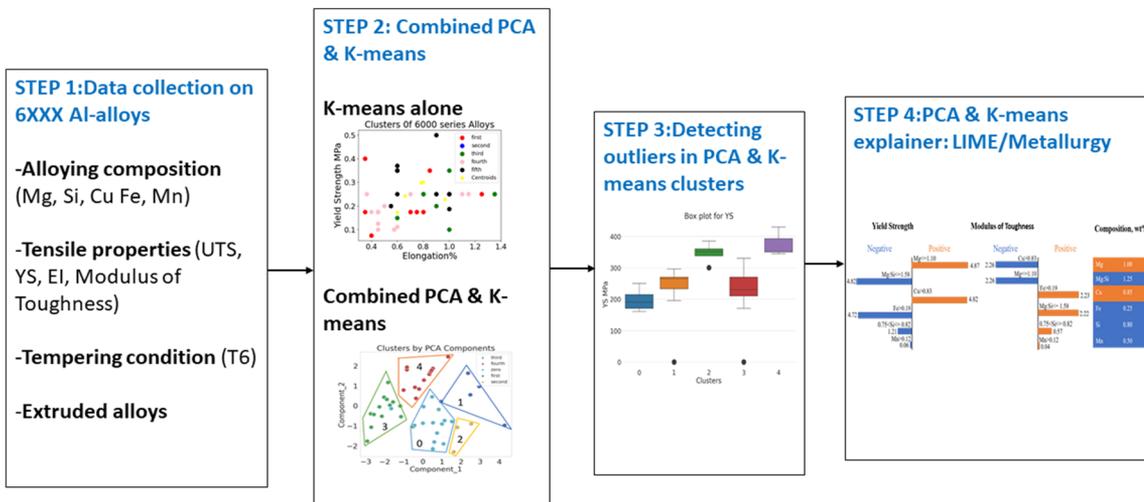


Fig. 1. Design loop.

Step 1: Standardize the data: If in the dataset the variables were measured on different value scales, standardization is performed.

Step 2: Perform PCA: PCA was applied to the standardized data to decrease the dataset dimensionality and determine the principle components.

Step 3: Choose the number of principle components: Based on the amount of variation (by a variance plot), we decided how many principle components to keep.

Step 4: Transform data: Using the selected principle components, the original data were transformed into the reduced dimensional space.

Step 5: Decide the number of clusters: An elbow method was applied to the transformed data.

Step 6: Perform k-means clustering: K-means clustering was applied to assign each data point to the closest cluster centroid.

Step 7: Evaluate the results: The quality of clustering was evaluated by a silhouette score.

Step 8: Interpret the results: The clustering results were interpreted based on the features of the clusters, the features that were making the differences between the clusters and any other information about the dataset.

Step 9: Visualize the results: The clustering results were visualized.

Box Plot to Detect Outliers in Clusters

A box plot is a graphical representation to detect outliers for each cluster based on the distribution of variables in each cluster. For each cluster, it consists of a whisker and box plot with whiskers representing the range of the data and the box representing the interquartile range (IQR) of the data points. Outliers are plotted beyond the whiskers as individual points. We visualized and inspected the distribution of the alloy features in each cluster by box plot and identified any potential outliers present in them. Any point outside the whiskers of the box plot was identified as an outlier. By interpreting these plots for each cluster, we could determine which clusters had outliers in high proportion and which clusters had more homogeneous distributions or no outliers.⁴⁵

LIME Algorithm

The combined PCA and K-means clustering was applied to visualize clear clusters and box plot to detect outliers. The next step was to perform LIME algorithm to understand how the compositions were affecting properties in each cluster. For LIME, we ignored the outliers.⁴⁶ The main focus of LIME is to train local surrogate models for explaining individual clusters predicted by PCA & K-means algorithm.⁴⁷ To explain the clusters, the algorithm processes the input data samples, i.e., alloy compositions, and ranks the positive or negative impacts of compositions on tensile properties in each cluster.⁴⁶ The LIME algorithm can also be used in

regression mode to explain the predictions of regression models. In this way we can explain the meaning behind clustering to get better understanding of how properties are varying with compositions in our black box clustering model.

RESULTS AND DISCUSSION

Data Collection

It is necessary to have access to high-quality data to accurately apply the machine learning algorithms. Fifty sets of 6XXX series Al alloys were collected from the ASM Specialty Handbook,¹ makeitlearn website⁴⁸ and literature search, and presented as supplementary material; see the supplementary data file. The dataset consisted of eight dimensions of explanatory variables including composition, yield strength (YS) and modulus of toughness. It is well known that the volume fraction, morphology and type of precipitates as well as the condition of the alloys (deformed, recrystallized, etc.) define the properties in Al alloys. Therefore, for the data consistency and accuracy of ML predictions, and to reduce the variation in alloy characteristics caused in by various processing conditions, the alloy properties collected were limited to the T6 tempering condition of extruded alloys with the main focus on tensile properties.

The alloying elements were within a composition space of $0.05 < \text{Cu} < 0.95$, $0 < \text{Mn} < 1.1$, $0.35 < \text{Mg} < 1.4$, $0 < \text{Zn} < 0.75$, $0.35 < \text{Si} < 1.35$ and $0.075 < \text{Fe} < 0.4$. We excluded minor alloying elements such as Ti, Cr and V from our dataset as they did not have much effect on T6 tensile properties. In the present work, we only considered the tensile properties, i.e., the yield strength and modulus of toughness, as the main criteria to cluster the 6XXX series alloy grades. These properties are mainly affected by the addition of major alloying elements/impurities such as Cu, Si, Mg, Fe and Mn and very little by the addition/presence of minor elements or impurities. Small amounts of transition elements such as Cr, V and Ti mainly affect the electrical resistivity and are used to control recrystallization.¹ Corresponding to this composition space and the T6 tempering condition, the tensile properties, i.e., tensile strength, yield strength and elongation, were collected, and the modulus of toughness was calculated as $0.5 \times (\text{UTS} + \text{YS}) \times \text{EI}$.⁴⁹ We used the modulus of toughness to reduce the number of variables in the machine learning model. In future, we will extend our dataset to other properties. The elongation (EI), yield strength (YS), tensile strength (UTS) and modulus of toughness were in the ranges of $3 < \text{EI} < 15$, $170 \text{ MPa} < \text{YS} < 430 \text{ Mpa}$, $210 \text{ Mpa} < \text{UTS} < 483 \text{ Mpa}$ and $10.5 < \text{modulus of toughness} < 49$. As for the precipitation hardening response, the Mg:Si ratio is important in the 6XXX series Al alloys; it was calculated and found to be in the range of $0.34 < \text{ratio} < 1.71$ in the dataset. The

ratio of magnesium (Mg) to silicon (Si), most commonly referred to as the Mg:Si ratio, is important because of its influence on the alloy's precipitation hardening characteristics and kinetics, affecting the mechanical properties.⁵⁰

Combined Clustering and PCA

To investigate similarities between the aluminum alloys in our dataset in terms of composition and tensile properties, specifically yield strength and modulus of toughness, it was necessary to perform unsupervised learning, i.e., clustering analysis, to search for clusters of aluminum alloys having similar property and compositional range. To reveal clusters in the dataset, we performed k-means clustering in conjugation with PCA. This allowed us to obtain rather well-defined clusters.

Figure 2 shows the raw data set with two selected dependent parameters, i.e., yield strength and modulus of toughness. All points in our current data set are represented in this figure. These data points show that the domain for the modulus of toughness is from 10.5 MPa to 49 MPa, whereas for YS it is from 170 MPa to over 430 MPa, which indicated a significant contrast in the value ranges. Therefore, standardization of data was applied. Standardization was an important part of data pre-processing. Generally, we want to treat each feature equally. To keep the variations between the feature values equivalent, we can achieve this by transforming the features so that their values lie within the same numerical range. In the next step, we reduced the number of features in our data set by a dimensionality reduction technique, i.e., cumulative variance plot. Y-axis in Fig. 3 shows the amount of variance captured depending on the number of principle components as shown on X-axis. In the PCA variance plot, the number of principle components indicates how many principle components are required to explain a significant proportion of the total variance in the dataset. Each

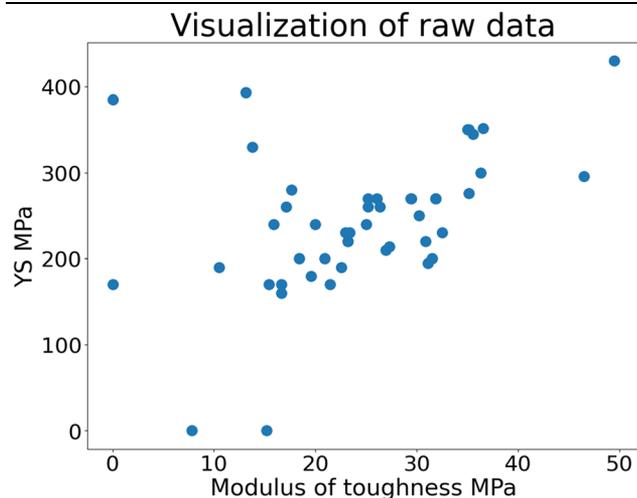


Fig. 2. Scatter plot YS versus modulus of toughness.

principle component captures a certain amount of variability in the original dataset. The higher the number of principle components, the more variance in the data is accounted for. The trade-off between the amount of variance explained and the number of principle components is understood using this plot. It helps in dimensionality reduction or better visualization by determining the optimal number of principle components to retain for further analysis.⁵¹

The plot often exhibits a declining pattern, with the first few principle components explaining a significant amount of variance while subsequent components explain less and less. When determining the number of principle components to retain, the elbow point or the point of inflection in the plot is often considered. It represents the point of diminishing returns, where including additional principle components provides diminishing increases in the explained variance. By analyzing the PCA variance plot, based on the desired amount of variance explained, we can make an informed decision about the number of principle components to select and the dimensionality reduction requirements for data analysis.

As a rule of thumb, at least 80% variance should be present in the principle component.⁵² So, in present case, we decided to choose three principle components with > 85% variance.

In the third step, with the chosen number of components, we performed PCA. We used PCA for better visualization of clusters. Now, for the elements in our data set we needed only the calculated resulting components scores that were used in the k-means algorithm. Before going further with the k-means algorithm, we implemented an elbow method to find the total number of clusters present in our 6XXX series dataset considering the PCA scores. Figure 4 suggests the number of clusters to keep.

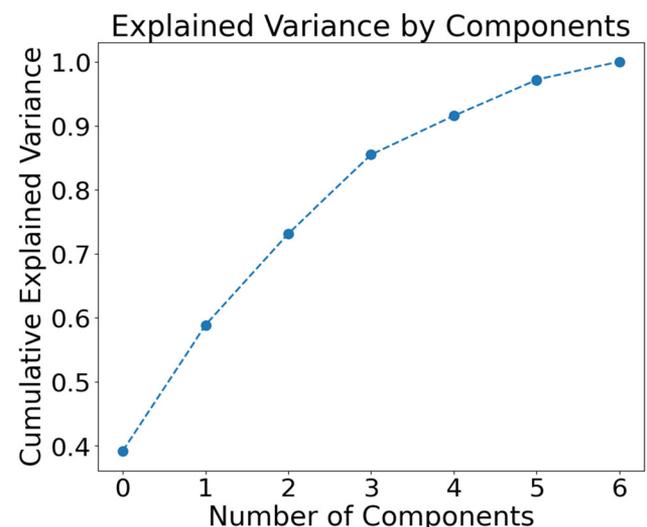


Fig. 3. Cumulative variance versus number of principle components plots.

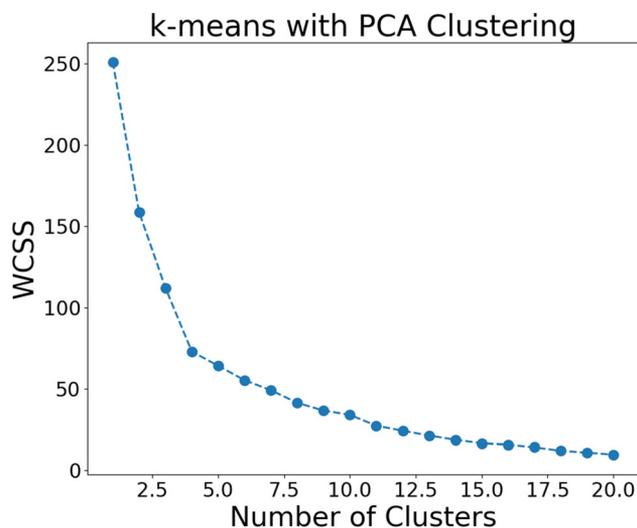


Fig. 4. Elbow plot to detect number of clusters.

The approach was based on finding a kink or “elbow” in the weighted cumulative sum of square (WCSS) distance curve.⁴⁴ In general, the part of the curve before the elbow would be rapidly decreasing, while the part after it would be going more gradually. In our case, the kink came between the four and five cluster mark. After trying both values, we decided to keep five clusters of alloys for quality visualization.

Figure 5 demonstrates the advantages of using k-means together with PCA in plot (a). It shows rather distinct clusters based on combined PCA and K-means algorithm compared with using K-means alone (Fig. 5b).

This was the main objective of using PCA prior to K-means clustering—to make fewer variables by combining them into more significant principle components for better visualization. Also, difference between the components is as large as possible, i.e., they are ‘orthogonal’ to each other. There were still some overlaps between clusters of 6XXX series Al alloys. The areas where the clusters overlapped were determined by the third principle component, which has low variance compared to other components not shown on this 2D graph.

We placed the cluster numbers in front of each set of data points (original features that comprised them); see supplementary data file. Based on the alloy property ranges and composition values in each cluster, we found the clusters are quite distinguishable with few overlaps in terms of composition and property values as given in Table I.

The ranking of concentration of critical alloying elements and YS is as follows in ascending order for all clusters:

$$\begin{aligned} \text{YS} &= 0 < 3 < 1 < 2 < 4, & \text{Si} &= 0 < 1 < 4 < 3 < 2, \\ \text{Cu} &= 0 < 3 < 1 < 2 < 4, & \text{Mg} &= 0 < 2 < 1 < 3 < 4, \\ \text{Mn} &= 0 < 1 < 4 < 3 < 2 \end{aligned}$$

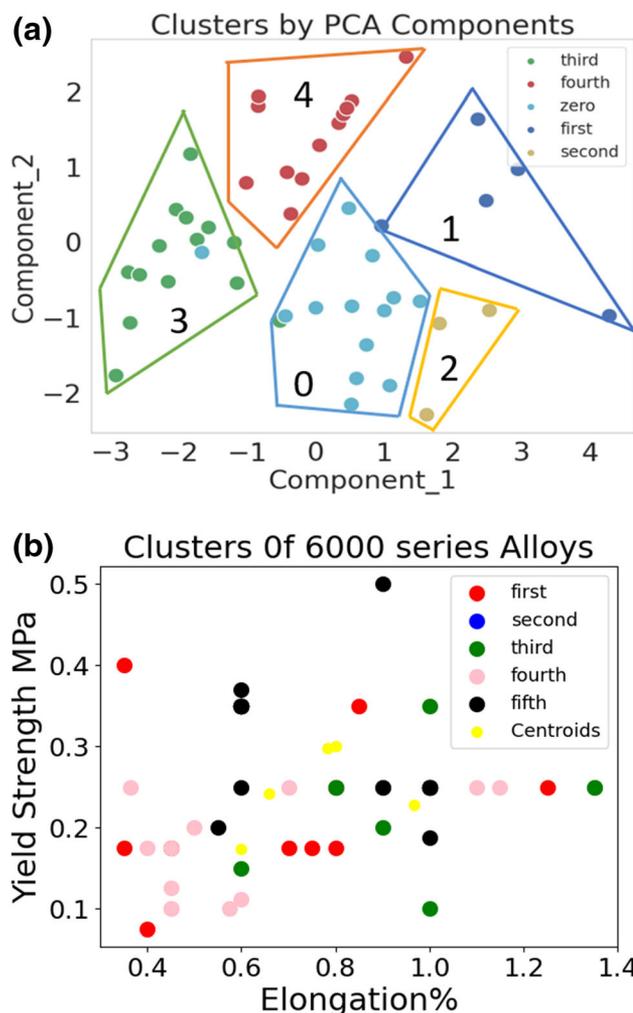


Fig. 5. K-means clustering with PCA (a) and without PCA (b).

This is also summarized in Table II with some example alloy grades in each cluster. Table II works in parallel with Table I or represents a summary table, from which the observation is made on how the concentration of alloying elements leads to low, medium and high values of the given properties. Table II gives us the base for metallurgical reasoning behind the clusters obtained from PCA and K-means algorithm; see Section “[Metallurgy-Based Understanding of Clusters](#)”.

It seems that in Cluster 4 containing alloys with large amounts of solute elements, i.e., Mg + Si + Cu, high Cu content, Si excess and medium levels of iron, the alloys demonstrated the highest YS. Cluster 1, because there are medium levels of solute elements, i.e., Mg + Si + Cu \sim 1.9, Cu content is medium, and Mg:Si is in stoichiometry, comprised the alloys with medium levels of YS values. Its modulus of toughness (see Table I) is higher though than in Cluster 3. Alloys grouped in Cluster 0 showed the lowest YS due to the low percentage of solute elements, i.e., Mg + Si + Cu \sim 1.2, low Cu content and only slight Si excess. Alloys in Cluster 2

Table I. Composition and property value ranges in the five clusters during ML

Group	0	1	2	3	4
Si	0.37–0.45	0.5–0.6	1–1.35	1–1.25	0.8–0.9
Fe	0.1–0.18	0.35–0.37	0.25	0.25–0.35	0.25
Cu	0.05–0.1	0.2–0.28	0.28–0.95	0.1–0.2	0.85
Mn	0.03–0.13	0.15	0.7–0.85	0.6–0.7	0.5–0.6
Mg	0.45–0.65	0.9–1	0.85–1.1	0.8–0.9	1–1.4
YS	170–220	220–280	300–385	220–270	350–430
Modulus of toughness	10.5–22	25–35	35.15–36.55	22–27	35.03–49
Mg:Si	1.3–1.68	1.667	< 1	< 1	1–1.667
Cu content	Low	Medium	High	Low	High
Excess Si	Slight excess	Stoichiometry	Large excess	Large excess	Excess
Mg + Si + Cu	1.2 (low)	1.9 (medium)	2.9 (high)	2.1 (medium)	2.65 (high)

Table II. Composition values affecting yield stress and modulus of toughness

Clusters	Composition variation in each cluster	Alloy grades
Cluster 4 (the highest YS)	High Cu, Mg and medium Si, Fe, Mn concentrations	6110A, 6013, 6092
Cluster 2 (2nd highest YS)	High Si, Cu, Mn, low Fe, Mg	6024, 6070, 6066
Cluster 1 (medium YS)	Low Mn, Si, high Fe, medium Cu, Mg	6151, 6053, 6061
Cluster 3 (2nd lowest YS)	High Si, Fe, Mg, Mn, low Cu	6005A, 6009, 6012
Cluster 0 (lowest YS)	Low Fe, Si, Mg, Mn, and Cu	6106, 6101A, 6060

had a high percentage of solute elements, i.e., $Mg + Si + Cu \sim 2.9$, and large excess of Si, but the percentage of Cu and Mg was less than in Cluster 4. Therefore, these alloys showed the second highest YS. Cluster 3 included alloys with medium levels of solute elements, i.e., $Mg + Si + Cu \sim 2.1$, low Cu and high Fe content, $Mg:Si \sim 0.9$, which was reflected in a relatively low YS and modulus of toughness.

Box Plot Analysis

A box plot for each cluster based on the YS and modulus of toughness shows the distribution of values as well as the potential outliers. We checked that some data points are legitimate data points and some are errors or anomalies that need to be removed or corrected. Some box plots (Fig. 6a and b) are very compressed because the data are highly concentrated within a narrow range of YS and modulus of toughness values. This may indicate that there was very little variation in the data point values, i.e., YS and modulus of toughness values within that cluster were tightly grouped around a central value, meaning that there was little deviation from the mean or median value.⁴⁵ The box plots in Fig. 6a and b clearly show that Clusters 0, 1, 2 and 3 had missing values of YS and modulus of toughness (0 values), which might be due to some missing values of El and UTS in the data set. Clusters 1 and 4 showed outliers with very high modulus of toughness values as well as with much

lower values. No outliers in terms of YS were found for Clusters 0 and 4.

In the next step (Section “LIME Algorithm”), we ignored these outliers to perform LIME algorithm for most common alloy composition for each of the clusters to understand which alloying elements had positive impact and which had negative impact on YS and modulus of toughness to understand the physics and patterns inside the clustering. After removing the outliers, we received much better defined clusters. Later, in Section “Metallurgy-Based Understanding of Clusters”, we considered these outliers to connect the metallurgical reasoning behind the clusters obtained from combined approach of PCA and K-means. In this work our main aim is to narrow down the search space of aluminum alloys to a few groups, which will be the basis for future designing of 2–3 alloys for each cluster.

LIME Algorithm

As the model becomes more complex, the prediction becomes more accurate; however, the interpretability becomes challenging. Therefore, clustering techniques are known as black box algorithms. To verify the model reliability, it is important to check how the model is operating. Therefore, an explainable artificial intelligence (XAI) emphasis should be placed on the chosen factors, or compositions, that have a significant impact on how the clusters are defined. A black box clustering model whose internal reasoning is hidden

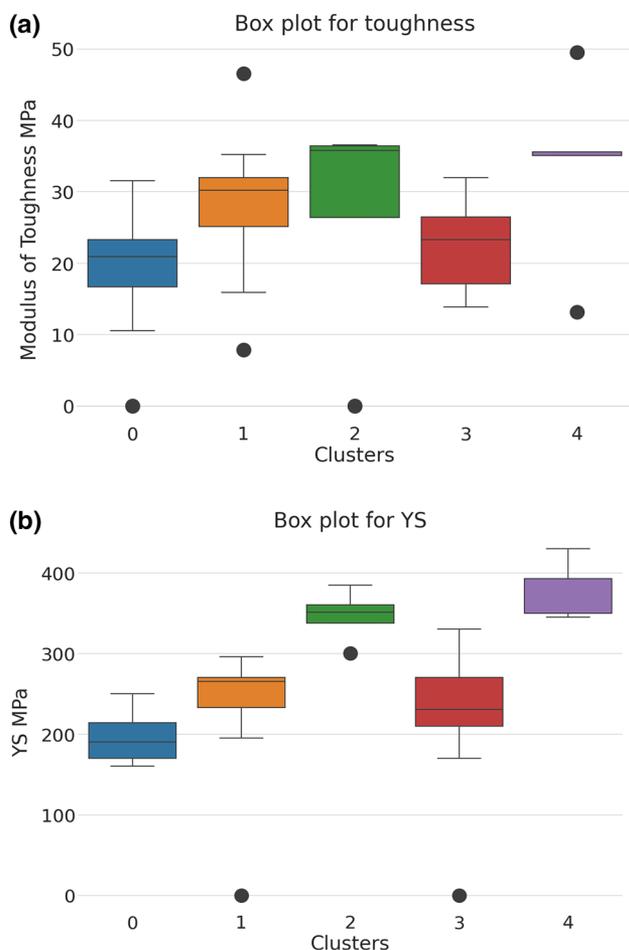


Fig. 6. Box plot to detect outliers in each cluster for (a) modulus of toughness and (b) YS.

and difficult to understand can be explained using the LIME algorithm. As described above, the clusters consisting of alloys with a varying range of properties and composition values were predicted using PCA and K-means combined algorithm. To understand and interpret the contribution of each alloying elements on the tensile properties, we needed to use an extra step, as clustering could not distinguish between independent and dependent variables. In addition, it is necessary to verify the predicted results with the scientific knowledge available. The clusters obtained from combined algorithm of PCA and K-means were explained here using the LIME algorithm. For each cluster, a separate dataset was prepared (see supplementary data file). The LIME algorithm divided the dataset consisting of alloy features into training and test sets and interpreted the prediction of target properties, i.e., YS and modulus of toughness, as shown in Fig. 7. It shows the alloying elements that positively or negatively affect the yield strength and modulus of toughness. For each cluster, we selected a representative alloy and interpreted its inclusion in the cluster. When considering the

effect, one also needs to consider the number of elements in the alloy.

For Cluster 0, all alloying elements, except Mn, had a negative impact on the YS values; consequently, the YS of the alloys in this cluster was the lowest. On the other hand, almost all alloying elements (except Si) should have had a positive impact on the modulus of toughness. However, due to the lower amount of these alloying elements in this cluster, this contribution could not be realized to the full extent, and the overall modulus of toughness values were low.

For Cluster 4, which showed the highest YS and moderate modulus of toughness, the alloys had high concentrations of Mg and Cu, and these elements affected the YS most, while decreasing the modulus of toughness. Fe had a negative effect on YS and positive on the modulus of toughness, but because of its low concentration these effects were not realized to the full extent.

Metallurgy-Based Understanding of Clusters

In the previous section, we looked at some reasoning behind the YS and modulus of toughness values obtained in each alloy clusters because of the varying percentage of solute elements present and Mg:Si ratio. This, however, was done using a ML LIME algorithm so was based formally on the numbers in the dataset. In this section, we will look at the metallurgical meaning behind the grouping of the alloys in each cluster, based on known mechanisms of precipitation hardening in 6XXX series alloys.

According to the Al-Mg-Si system phase diagram, the precipitation events in ternary 6XXX series alloys were supersaturated solid solution (SSS) \rightarrow GP \rightarrow β'' \rightarrow β' \rightarrow β .^{50,53–55} There is consensus that the metastable coherent β'' phase serves as the main strengthening phase in Al-Mg-Si ternary alloys.^{53–55}

When Cu is added to the Al-Mg-Si alloys in the 6XXX series, the Al-Mg-Si-Cu family of alloys is created. These quaternary alloys do not have a distinct designation within the 6XXX alloy series.^{50,56,57} The possibility for the formation of a quaternary phase is a significant underlying characteristic shared by all these alloys.⁵⁸ The phase was extensively studied and is so called the Q phase.^{59–62} While the compositions of ternary Al-Mg-Si alloys fall on the equilibrium phase diagram at normal aging temperatures in a three-phase field, i.e., (Al) + β (Mg₂Si) + (Si),⁵⁸ the coexisting equilibrium three-phase fields enlarge into three tetrahedron composition spaces upon the inclusion of Cu. Inside each of these spaces, a four-phase equilibrium is present in the equilibrium phase diagram consisting of the two common phases, i.e., quaternary Q phase and (Al), and two other phases from the selection: β (Mg₂Si), θ (CuAl₂) and (Si).⁵⁸ Regarding the precipitation, which is a

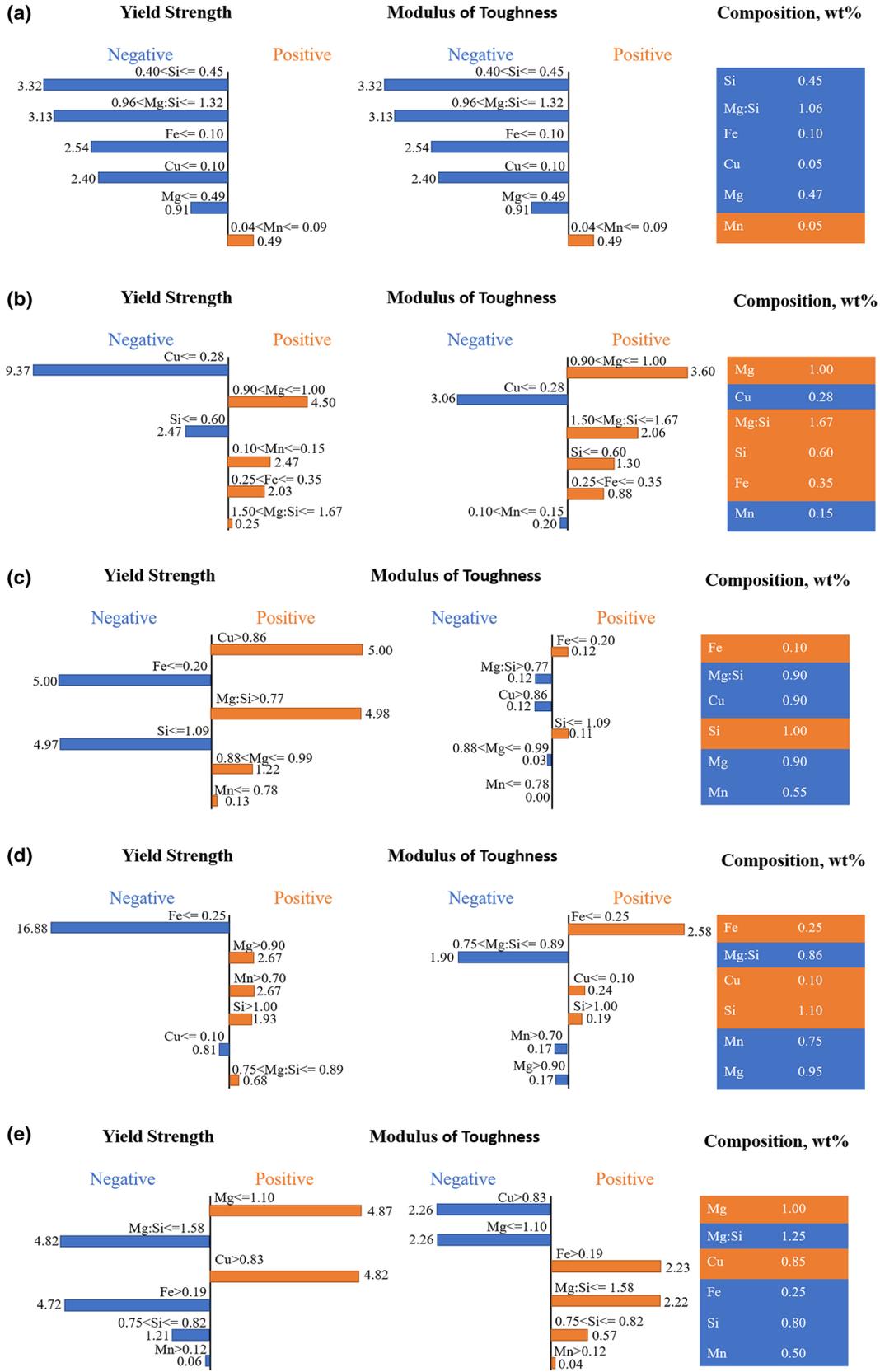


Fig. 7. LIME plots of YS and modulus of toughness: (a) Cluster 0 (4th data), (b) Cluster 1 (31th data), (c) Cluster 2 (27th data), (d) Cluster 3 (49th data), (e) Cluster 4 (13th data).

metastable process, the presence of Cu modifies the order of precipitation as follows.⁶³

At a lower Cu concentration: SSS—GP zones— β'' (with Cu?)—Si—numerous variations of β' (with Cu) including $\beta'C$ — Mg_2Si and Si.

At a higher Cu concentration: SSS—GP zones— β'' (with Cu?)— θ' or/and Q' —Si—numerous variations of β' (with Cu) including $\beta'C$, Q' and θ' — Q (AlMgSiCu), Mg_2Si , Al_2Cu and Si.

These two precipitation paths may form at intermediate copper concentrations in cases of natural aging preceding, artificial aging and temperatures $> 200^\circ C$. It is also known that the excess of Si regarding Mg_2Si stoichiometry (revealed by the Mg:Si ratio) is beneficial to hardening because of the effect on the composition and structure of GP zones. The addition of copper also makes 6XXX series alloys less sensitive to the negative effects of natural aging on hardening upon artificial aging.⁶³

Based on this general understanding of metallurgy of 6XXX series alloys, we can suggest the metallurgical reasons behind grouping of the alloys in different clusters. Let us look primarily at the yield strength as it is the property that responds most directly to the precipitation hardening.

Cluster 4: Cu = 0.85%, high concentration of Mg and Si, excess of Si (Mg:Si = 1–1.67), low Fe. High YS is due to the formation of β'' modified with Cu.

Cluster 2: Medium range of Cu = 0.275–0.95%, higher concentration of Mg and Si, large excess of Si (Mg:Si < 1). High YS due to the formation of β'' modified with Cu.

Cluster 1: Cu = 0.2–0.275%, medium Si and high Mg concentration, Mg:Si is close to stoichiometry ~ 1.667 , high Fe. Medium YS due to the less β'' particles and high Fe that takes some Si from the solid solution.

Cluster 3: Cu = 0.1–0.2%, high concentration of Mg and Si and large excess of Si (Mg:Si < 1), high Fe. Medium YS due to the precipitation of β'' but in lesser quantities due to the consumption of Si by Fe-containing phases.

Cluster 0: Cu = 0.05–0.1%, lower concentration of Mg and Si and moderate excess of Si (Mg:Si = 1.3–1.68). The lowest YS due to the overall smaller concentration of solute elements and the resultant smaller amount of the hardening β'' phase.

These observations proved that clustering had a metallurgical reasoning behind it as the variation in the tensile properties between the clusters was due to the various precipitation hardening features. These clusters will now be the basis for future optimization and reduction of the number of the alloy grades without compromising their tensile properties.

CONCLUSION

In the present study, we attempted to refine the search space of 6XXX series aluminum alloys (extruded, T6 condition) into a few clusters by a

combined PCA and K-means algorithm. We successfully created five clusters, each having distinct ranges of composition and properties with very few overlaps and outliers. To find the characteristic features that make each cluster special, we used a LIME algorithm. This allowed us to identify the alloying elements with their combinations and concentrations that affected the yield strength and modulus of toughness of the alloys in each cluster.

In addition to this formal selection of the features determining the formation of each cluster, we also examined the alloys in the clusters from the point of view of the known metallurgical knowledge due to precipitation hardening in 6XXX series alloys, as precipitation hardening of 6XXX alloys has a direct impact on the yield strength. This analysis showed that the amount of alloying elements, Mg:Si ratio and the contents of Fe are the parameters that determined the inclusion of the alloys in a given cluster.

This study showed that the machine learning algorithms gave us a meaningful selection of five clusters that incorporated 50 commercial 6XXX series alloy grades. This is a good base for optimization and reducing the number of the alloy grades without compromising their tensile properties. In the future, we will widen the selection of properties in our dataset.

SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1007/s11837-023-06025-9>.

ACKNOWLEDGEMENTS

This work was done within the framework of Circular Metals Centre funded by an UKRI/EPSC grant EP/V011804/1. The authors thank Prof. H. Assadi and I. Chang for fruitful discussions. The corresponding author acknowledges the financial support from Brunel University London for the scholarship.

DATA AVAILABILITY

The raw/processed data required to obtain these results can be shared upon reasonable request to the corresponding author.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could influence this work.

OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

REFERENCES

- J.R. Davis (Ed.) *Aluminum and Aluminum Alloys* (ASM International, Materials Park, OH, 1993), pp. 3–88.
- D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, *Nat. Commun.* <https://doi.org/10.1038/ncomms11241> (2016).
- L.F. Mondolfo, *Aluminium Alloys: Structure and Properties* (Butterworths, London, 1979), pp.787–797.
- B. Ajayi, S. Kumari, D. Jaramillo-Cabanzo, J. Spurgeon, J. Jasinski, and M. Sunkara, *J. Mater. Res.* <https://doi.org/10.1557/jmr.2016.92> (2016).
- S. Curtarolo, G. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, *Nat. Mater.* <https://doi.org/10.1038/nmat3568> (2013).
- Y.J. Soofi, M.A. Rahman, Y. Gu, and J. Liu, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2022.111783> (2022).
- J.R. Duflou, A.E. Tekkaya, M. Haase, T. Welo, K. Vanmeensel, K. Kellens, W. Dewulf, and D. Paraskevas, *CIRP Ann.* <https://doi.org/10.1016/j.cirp.2015.04.051> (2015).
- J. Gronostajski, H. Marciniak, and A. Matuszak, *J. Mater. Process. Technol.* [https://doi.org/10.1016/S0924-0136\(00\)00634-8](https://doi.org/10.1016/S0924-0136(00)00634-8) (2000).
- T. Alam and A.H. Ansari, *Int. J. Adv. Technol. Eng. Sci.* 5(5), 278 (2017).
- V. Kevorkian, *Mater. Technol.* 47(1), 13 (2013).
- W.Y. Wang, J. Li, W. Liu, and Z.K. Liu, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2018.11.001> (2019).
- A. Asatiani, P. Malo, P.R. Nagbøl, E. Penttinen, T. Rinta-Kahila, and A. Salovaara, *J. Assoc. Inf. Syst.* 22, 325 <https://doi.org/10.17705/1jais.00664> (2021).
- G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, *Rev. Mod. Phys.* <https://doi.org/10.1103/RevModPhys.91.045002> (2019).
- J. Wei, X. Chu, X.Y. Sun, K. Xu, H.X. Deng, J. Chen, Z. Wei, and M. Lei, *Info. Mat.* <https://doi.org/10.1002/inf2.12028> (2019).
- L. Himanen, M.O. Jäger, E.V. Morooka, F.F. Canova, Y.S. Ranawat, D.Z. Gao, P. Rinke, and A.S. Foster, *Comput. Phys. Commun.* <https://doi.org/10.1016/j.cpc.2019.106949> (2020).
- P. Raccuglia, K.C. Elbert, P.D. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, and A.J. Norquist, *Nature.* <https://doi.org/10.1038/nature17439> (2016).
- C. Wang, H. Fu, L. Jiang, D. Xue, and J. Xie, *npj Comput. Mater.* <https://doi.org/10.1038/s41524-019-0227-7> (2019).
- P.V. Balachandran, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2019.03.057> (2019).
- D. Xue, D. Xue, R. Yuan, Y. Zhou, P.V. Balachandran, X. Ding, J. Sun, and T. Lookman, *Acta Mater.* <https://doi.org/10.1016/j.actamat.2016.12.009> (2017).
- V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, *npj Comput. Mater.* <https://doi.org/10.1038/s41524-018-0085-8> (2018).
- J. Gao, Y. Liu, Y. Wang, X. Hu, W. Yan, X. Ke, L. Zhong, and Y. He, *Ren. Phys. Chem. C.* <https://doi.org/10.1021/acs.jpcc.7b04636> (2017).
- R. Yuan, Z. Liu, P.V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, and T. Lookman, *Adv. Mater.* <https://doi.org/10.1002/adma.201702884> (2018).
- C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, and Y. Su, *Acta Mater.* <https://doi.org/10.1016/j.actamat.2019.03.010> (2019).
- W. Xu, P.R. del Castillo, and S. Van Der Zwaag, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2008.11.006> (2009).
- Y. Zeng, Q. Li, and K. Bai, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2017.12.030> (2018).
- H. Wu, A. Lorensen, B. Anderson, L. Witteman, H. Wu, B. Meredig, and D. Morgan, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2017.03.052> (2017).
- X. Jiang, H.Q. Yin, C. Zhang, R.J. Zhang, K.Q. Zhang, Z.H. Deng, G. Liu, and X. Qu, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2017.09.061> (2018).
- Z. Deng, H. Yin, X. Jiang, C. Zhang, K. Zhang, T. Zhang, B. Xu, Q. Zheng, and X. Qu, *JOM.* <https://doi.org/10.1016/j.jom.2018.07.049> (2018).
- B. Wang, W. Zhao, Y. Du, G. Zhang, and Y. Yang, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2016.08.035> (2016).
- A. Rovinelli, M.D. Sangid, H. Proudhon, and W. Ludwig, *npj Comput. Mater.* <https://doi.org/10.1038/s41524-018-0094-7> (2018).
- A.O. Furmanchuk, J.E. Saal, J.W. Doak, G.B. Olson, A. Choudhary, and A. Agrawal, *J. Comput. Chem.* <https://doi.org/10.1002/jcc.25067> (2018).
- T.D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, *npj Comput. Mater.* <https://doi.org/10.1038/s41524-017-0042-y> (2017).
- Y.T. Sun, H.Y. Bai, M.Z. Li, and W.H. Wang, *J. Phys. Chem. Lett.* <https://doi.org/10.1021/acs.jpclett.7b01046> (2017).
- J. Wang, X. Yang, Z. Zeng, X. Zhang, X. Zhao, and Z. Wang, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2017.06.015> (2017).
- S. Chmiela, A. Tkatchenko, H.E. Sauceda, I. Poltavsky, K.T. Schütt, and K.R. Müller, *Sci. Adv.* <https://doi.org/10.1126/sciadv.1603015> (2017).
- M.S. Ozerdem and S. Kolukisa, *Mater. Des.* <https://doi.org/10.1016/j.matdes.2008.05.019> (2009).
- V. Revi, S. Kasodariya, A. Talapatra, G. Pilania, and A. Alankar, *Comput. Mater. Sci.* <https://doi.org/10.1016/j.commatsci.2021.110671> (2021).
- G. Peng, Y. Cheng, Y. Zhang, J. Shao, H. Wang, and W. Shen, *J. Manuf. Syst.* <https://doi.org/10.1016/j.jmsy.2022.08.014> (2022).
- S.M. Golovanov, V.I. Orlov, L.A. Kazakovtsev, and A.M. Popov, *IOP Conf. Ser.: Mater. Sci. Eng.* <https://doi.org/10.1088/1757-899X/537/2/022035> (2019).
- A.L. Kazakovtsev, A.N. Antamoshkin, and V.V. Fedosov, *IOP Conf. Ser.: Mater. Sci. Eng.* 122, 012011 <https://doi.org/10.1088/1757-899X/122/1/012011> (2016).
- F. Bießmann, T. Rukat, P. Schmidt, P. Naidu, S. Schelter, A. Taptunov, D. Lange, and D. Salinas, *J. Mach. Learn. Res.* 20(175), 1 (2019).
- V. Kevorkian, *Metallurgia* 16, 103 (2010).
- C. Ding and X. He, in *Proceedings of the Twenty-First International Conference on Machine Learning* (2004). <https://doi.org/10.1145/1015330.1015408>.
- J.A. Hartigan and M.A. Wong, *Royal Stat. Soc. Ser. C Appl. Stat.* <https://doi.org/10.2307/2346830> (1979).
- D.F. Williamson, R.A. Parker, and J.S. Kendrick, *Ann. Intern. Med.* <https://doi.org/10.7326/0003-4819-110-11-916> (1989).

46. M.W. Craven and J.W. Shavlik, *Adv. Neural Inf. Process. Syst.* p. 8 (1995).
47. D. Garreau and U. Luxburg, in: *International Conference on Artificial Intelligence and Statistics* (2020), pp. 1287–1296. <https://doi.org/10.48550/arXiv.2001.03447>.
48. Make it from, “Aluminum Alloys-Materials-Engineering,” (2022) <https://www.makeitfrom.com/material-group/Aluminum-Alloy>. Accessed 18 Apr 2022.
49. H. Kuhn (Ed.) *ASM Handbook*, vol. 8 (ASM International, Mechanical Testing and Evaluation, Materials Park (OH), 2000). <https://doi.org/10.31399/asm.hb.v08.9781627081764>.
50. D.J. Chakrabarti and D.E. Laughlin, *Prog. Mater. Sci.* [https://doi.org/10.1016/S0079-6425\(03\)00031-8](https://doi.org/10.1016/S0079-6425(03)00031-8) (2004).
51. H. Abdi and L.J. Williams, *Comput. Stat.* <https://doi.org/10.1002/wics.101> (2010).
52. W.F. Miao and D.E. Laughlin, *Scr. Mater.* [https://doi.org/10.1016/S1359-6462\(99\)00046-9](https://doi.org/10.1016/S1359-6462(99)00046-9) (1999).
53. G.A. Edwards, K. Stiller, G.L. Dunlop, and M.J. Couper, *Acta Mater.* [https://doi.org/10.1016/S1359-6454\(98\)00059-7](https://doi.org/10.1016/S1359-6454(98)00059-7) (1998).
54. A.K. Gupta, D.J. Lloyd, and S.A. Court, *Mater. Sci. Eng A.* [https://doi.org/10.1016/S0921-5093\(00\)01814-1](https://doi.org/10.1016/S0921-5093(00)01814-1) (2001).
55. C.D. Marioara, S.J. Andersen, T.N. Stene, H. Hasting, J. Walmsley, A.T. Van Helvoort, and R. Holmestad, *Philos. Mag.* <https://doi.org/10.1080/14786430701287377> (2007).
56. The Aluminum Association. *International Alloy Designations and Chemical Composition Limits for Wrought Aluminum and Wrought Aluminum Alloys* (The Aluminum Association, Washington, 2001).
57. S.D. Dumolt, D.E. Laughlin, and J.C. Williams, *Scr. Mater.* [https://doi.org/10.1016/0036-9748\(84\)90362-4](https://doi.org/10.1016/0036-9748(84)90362-4) (1984).
58. M.W. Zandbergen, A. Cerezo, and G.D. Smith, *Acta Mater.* <https://doi.org/10.1016/j.actamat.2015.08.018> (2015).
59. A. Biswas, D.J. Siegel, and D.N. Seidman, *Acta Mater.* <https://doi.org/10.1016/j.actamat.2014.05.001> (2014).
60. C.S. Tsao, C.Y. Chen, U.S. Jeng, and T. Kuo, *Acta Mater.* <https://doi.org/10.1016/j.actamat.2006.06.005> (2006).
61. D.L.W. Collins, *J. Inst. Met.* **86**, 325 (1957–1958).
62. D.P. Smith, *Metallurgia* 63, 223 (1961).
63. D.G. Eskin, *J. Mater. Sci.* <https://doi.org/10.1023/A:1021109514892> (2003).

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.