Results in Materials 4 (2019) 100020

Contents lists available at ScienceDirect

Results in Materials

journal homepage: www.journals.elsevier.com/results-in-materials

Introducing self-organized maps (SOM) as a visualization tool for materials research and education

Jimin Qian [a,1], Nam Phuong Nguyen [a,1], Yutaka Oya [b], Gota Kikugawa [c], Tomonaga Okabe [a,b], Yue Huang [a,*], Fumio S. Ohuchi [a,**]

[a] Department of Materials Science and Engineering, University of Washington, Box 352120 Seattle, WA, 98195-2120, USA
[b] School of Engineering, Tohoku University, Aobayama, Aoba-ku, Sendai, 980-8579, Japan
[c] Institute of Fluid Science, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, 980-8577, Japan

ARTICLE INFO

Keywords:
Self-organizing map
Materials education
Ashby maps
Visualization tools

ABSTRACT

Materials informatics is an emerging discipline that opens a new paradigm of science: data-driven materials discovery. The visualization of high dimensional material properties data is challenging due to a lack of appropriate tools. This article describes a workflow for using Self-Organizing Maps (SOM) as a time- and cost-efficient method of visualizing and validating property relationships between materials. SOM is a machine learning technique that uses dimensionality reduction based on similarities between properties to visualize the relationships between materials in a high-dimensional dataset. While the conventionally used visualization method, Ashby plot, displays a pair of material's properties, SOM is a technique that can display multiple properties that are effective for high-dimensional materials correlation studies. In this paper, SOM was used to validate property correlations among 21 material properties in a dataset of over 398 commercial materials. We employed U-matrix map, clustering map and heatmaps to visualize the SOM we trained. The clustering and heat maps produced from SOM were used to identify unique materials and infer correlations between material properties and fundamental material structure. We have also shown that SOM method can provide multiple layers of interpretation through visualizing not only numerical properties but also categorical information of materials. Lastly, we have used SOM to provide new insight into the quantity of Grüneisen parameter across metals and ceramics. Through these examples, it is demonstrated that SOM can be used for different types of materials analyses for various applications. This paper advocates SOM as an integrated approach to study material property relationships that can be used for both materials education and discovery.

1. Introduction

In 2011, the Materials Genome Initiative (MGI) was initiated by the White House to find more time- and cost-effective methods for materials discovery, development, and manufacturing [1]. New sources of data from high-throughput experiments (both experimental and computational) are rapidly increasing in the fields of science and engineering, posing a significant difficulty in interpreting the high-dimensional data that is currently available [2]. For example, the National Institute of Standards and Technology (NIST) generates large volumes of property data that can offer insightful relationships between materials [3]. However, there are few tools available that are suited for visualizing and analyzing such complex high-dimensional data. Conventional methods of materials property comparison include Ashby map, which is a plot that displays two material properties at the same time. These plots are useful for determining ratios between different materials, but are not designed for visualizing multi-dimensional data.

Our project outlines a holistic method of approaching materials visualization through fundamental understanding of high-dimensional material relationships. Today, the heart of Materials Science and Engineering education is to integrate approaches of studying materials that include structure, characterization, processing, and performance factors of the materials. In 2008, the National Science Foundation (NSF) identified this task of integration in undergraduate materials science

* Corresponding author.
** Corresponding author.
E-mail addresses: huangyue@uw.edu (Y. Huang), ohuchi@uw.edu (F.S. Ohuchi).
[1] Authors share equal contribution to this project.

https://doi.org/10.1016/j.rinma.2019.100020
Received in revised form 23 August 2019; Accepted 29 August 2019
Available online 31 October 2019

curriculum as a key limitation to materials education and discovery [4]. Materials informatics is concerned with seeking high-throughput, robust methods to seek patterns among materials properties along different length scales. Even though informatics is established in fields such as biology and social sciences, materials informatics is growing in interest.

Our goals are to learn and reinforce existing knowledge about property-property relationships using a data science approach and an existing experimental data set of materials. In this paper, we introduce a Self-Organized Mapping (SOM) method to produce so-called clustering and heat maps that allow visual qualification of relationships between material properties. SOM uses experimental data to develop a powerful visualization tool that uses unsupervised machine learning to advance the knowledge of property-property relationships for materials informatics [5]. Due to its visual output, SOM is an effective and intuitive educational tool for teaching fundamentals of materials property relationships to a wide audience. We describe a comprehensive workflow that encompasses data management, machine learning, and finally, visualization to directly infer the relationship between materials. By using data visualization tools to analyze existing information, our goals are to validate the existing knowledge about property-property relationships of materials with a data science approach, to discover an efficient and practical way to understand the underlying relationships between materials that can potentially suggest new research directions. We have demonstrated that SOM can be used for different types of materials analyses such as using generated heat maps to evaluate positive and inverse correlations between properties. SOM also produces cluster maps which reveal several unique materials in our training dataset as a result of different processing methods and composite species. Lastly, we have used SOM to evaluate new quantities, such as Grüneisen parameter, that are not in our training data.

2. SELF-ORGANIZED mapping (SOM) visualization approach

2.1. SOM theory/Algorithm

In statistics, the dimension of data is defined as the number of variables that a data point has. In our study, the data is composed of many materials, and each data point represents a material with its mechanical, thermal, electrical and other properties. Thus, the number of dimensions our data contains corresponds to the number of properties of a material. Here, we introduce self-organizing map (SOM) as a powerful tool to visualize our multi-dimensional data of materials.

There are two types of machine learning: supervised and unsupervised. In supervised learning the main goal is to accurately predict the value of an outcome (regression) or categorize an outcome (classification) based on a number of input data. In unsupervised learning, the goal is to describe associations and patterns among a set of input data [6]. We used a self-organizing map (SOM) to do unsupervised machine learning to produce a low-dimensional, nonlinear approximation of a high-dimensional data, making it an appealing instrument for visualizing and exploring high-dimensional data [7]. In other words, SOM is employed to extrapolate the high-dimensional data set while retaining its topology [8]. The high-dimensional data that we want to approximate, or in other words, extrapolate, is called training data. Fig. 1 illustrates how a SOM is trained based on the training data, in which red dots 1 to 12 are the training data points in the multi-dimensional space, axes from a to f represent the variables of training data. In our case, the red dots 1–12 represent 12 materials, and the variables a to f are the properties of materials. Fig. 1(a) shows the initial status of the SOM. In this case, the SOM is a rectangular grid with the nodes (shown as black dots). Fig. 1(b) and (c) illustrate how the location of the grid and its nodes are updated. In Fig. 1(b), the nodes near the training data point 1 are 'dragged' to make them closer to it. In Fig. 1(c), the 'dragging' is operated on the

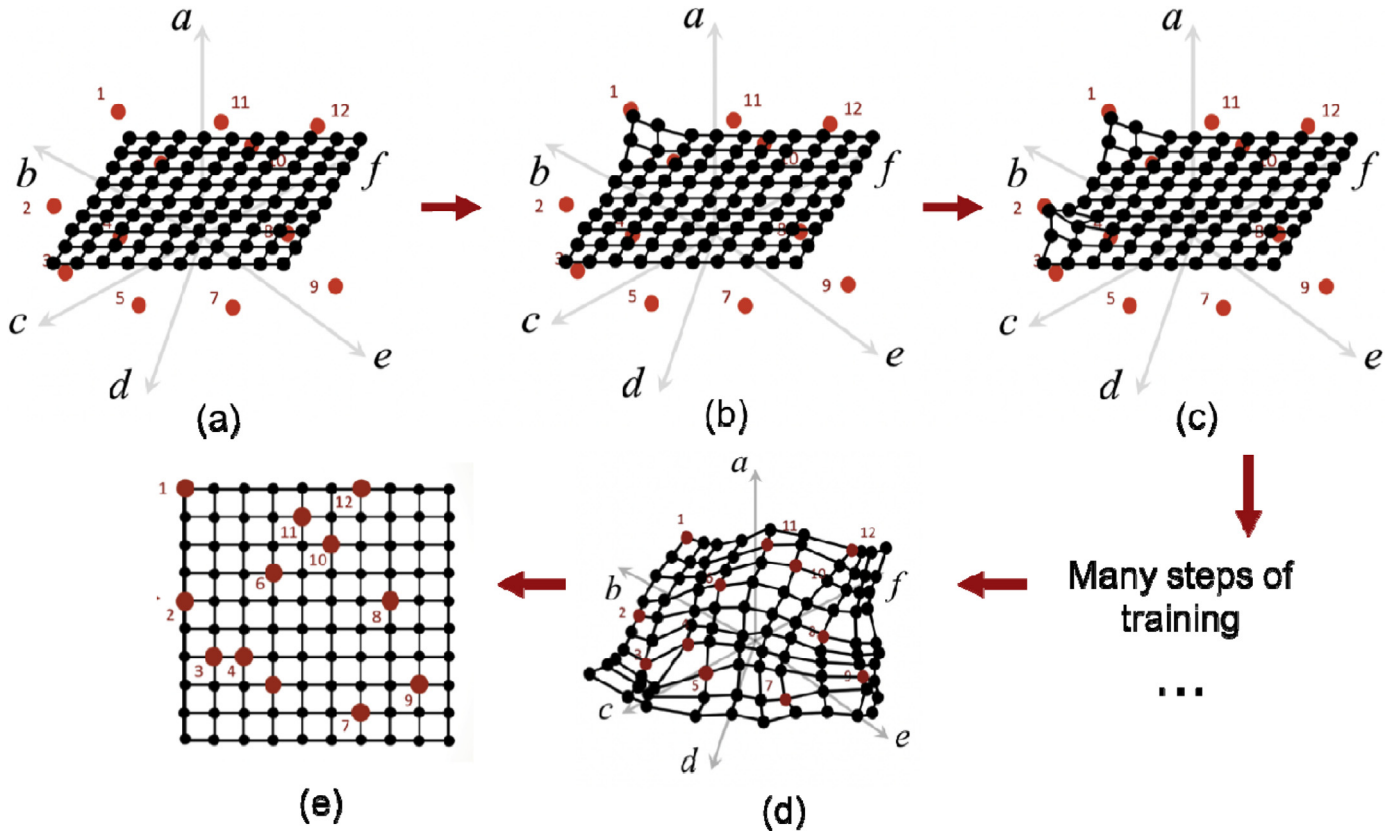

Fig. 1. Training of SOM (red dots are training data in the high-dimensional space, black dots and the grid are the trained SOM). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

nodes near the training data point 2. After many iterations of this 'dragging' operation, all the nodes get close enough to the training data and some nodes overlap with the training data points, as shown in Fig. 1(d). Although the grid is severely distorted in the multi-dimensional space during the training process, it retains its two-dimensional topology, which enables the visualization of the high-dimensional training data on a two-dimensional map. Fig. 1(e) shows the two-dimensional topology extracted from the grid in Fig. 1(d). The step from figure 1(d) to figure 1(e) is like flattening a crumpled paper, which finalizes the dimensionality reduction.

Among the many versions of SOM, we implemented a version of SOM called "*Batch Learning-SOM(BL-SOM)*" [9]. The algorithm of training the BL-SOM is briefly summarized as the following steps:

1) Define a topographic map in the multi-dimensional space spread by the training data (in our case, the map is a square grid with a given size), as shown in Fig. 1(a). For each node in the grid, its coordinate in the multi-dimensional space is called the weight of the node. The weights of all the nodes are initialized by a method called Principal Component Analysis at the beginning of training process [10].
2) A data point is chosen from the training data. For example in Fig. 2, the training data point 7 is chosen.
3) Calculate each node's Euclidean distance to the chosen point in step 2). Find the closest node to the chosen point in terms of the Euclidean distance in the multi-dimensional space. The node found in this step is called 'Best Matching Unit' (BMU) to the chosen training data point. For the example in Fig. 2, the best matching unit of training data point 7 is denoted as BMU7.
4) Find the neighbors of the BMU. The neighbors are the nodes within a given neighborhood radius. Note that the neighborhood radius refers to the distance on the 2D topology of the grid. As shown in Fig. 2(b), we assume the side length of each little block in the grid is of 1 unit, thus, the neighborhood radius here is 2. The radius of neighborhood is a value that starts large, then diminishes in each iteration.
5) Repeat step 2 to 4 to identify the BMU and neighbors of BMU for all the points in the training data.
6) Update the weights of all the BMUs as well as the weights of BMU's neighbors in a way that makes them closer to their corresponding training data point. One iteration is completed after this step.
7) Iterate until all the nodes in the map get close enough to the training data.

2.2. Current applications of SOM

In the engineering field, the applications of SOM widely range from process and system analysis, statistical pattern recognition, robotics, and telecommunications [11–14]. The most straightforward application of the SOM is in the identification and monitoring of complex machines and process states [15].

In materials science and engineering research, SOM is mainly applied to dimensionality reduction and visualization of high-dimensional materials data sets. In terms of dimensionality reduction, Principal Component Analysis (PCA) is actually the most popular tool and has been broadly used in materials informatics. Compared to PCA, SOM has advantage in maintaining the topological information of the training data and it is not inherently linear. Using PCA on multi-dimensional data can cause a loss of information when the dimension is reduced to two. If the target data has a lot of dimensions and each dimension is equally important, SOM can be quite effective over PCA. SOM is only one method of dimensionality reduction and visualization, with benefits over the standard PCA method depending on the research question and dataset [16]. Compared to an Ashby Map, which is a conventional way of visualizing materials property by separately displaying two properties on x and y axes in a coordinate system, SOM is another method of showing the relationships between materials and trade-off relationships in properties, especially when evaluating many properties at the same time [17]. For our materials training data with 21 properties, it will take 210 Ashby maps to display the relationship between every two properties. However, with the SOM method, only 1 SOM and 21 heatmaps (1 for each property) are necessary to intuitively display the relationships among properties, which will be discussed in later sections.

SOM is conventionally combined with other machine learning methods such as K-clustering and neural networks, which enable the interpretation of high-dimensional data in addition to visualization [18,

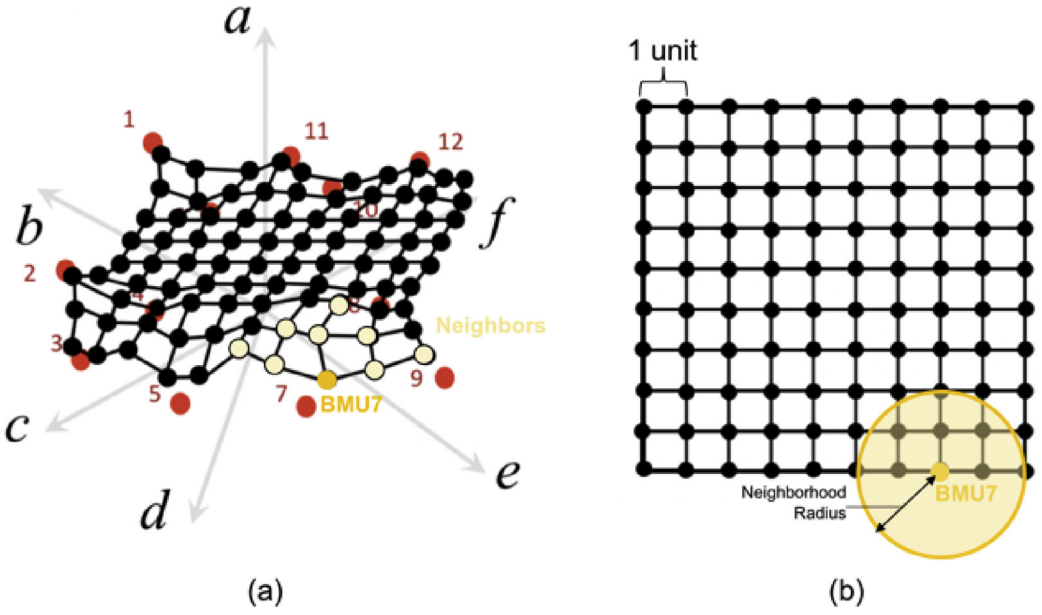

Fig. 2. Illustration of Best Matching Unit (BMU) and Neighborhood radius. 2(a) shows the SOM and training data in multi-dimensional space, and 2(b) shows the 2D topology of the SOM. 'BMU7' is the Best Matching Unit of training data point 7, the dots in light yellow are the neighbors of BMU7. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

19]. Kikugawa et al. used SOM and K-clustering method to classify consistent grouping of liquid substances in terms of their thermophysical properties [20]. Okabe's group at Tohoku University combined SOM with molecular dynamic simulation results to visualize the properties of complex structured polymers and discover the correlation between mechanical properties and multifunctional groups of base resins [21]. Rong Liu et al. implemented SOM-based clustering analysis and heat map visualization on high-throughput screening datasets for toxicity studies of engineered nanomaterials [22]. Schneider et al. employed SOM to perform feature extraction and dimensionality reduction on high-dimensional molecular descriptor vectors, and the SOM result is utilized as the input of neural network [23]. SOM is also applied to the prediction of a material's property. For example, Jha et al. used SOM to predict the composition of AlNiCo-type alloys that perform well for magnetic properties [24].

Overall, the application of SOM in materials science and engineering is so far limited, simply because of the lack of awareness of SOM's power in processing high-dimensional data. Additionally, the application of SOM is mostly focusing on standardized computational data, the research on commercial experimental data is also limited, due to the lack of consistency on experimental data measured by different researchers. However, in our paper, we demonstrated that the SOM method can be successfully applied to commercial experimental data from Granta, which will be shown in the next section.

In later sections, we will further discuss SOM's application in visualizing and analyzing relationships between material properties, which shows great potential in facilitating data-driven research in materials science field as well as materials selection in industry.

3. Methods

3.1. Data mining and extraction from Granta database

"CES" Selector is a commercially available database offered from Granta that contains MaterialUniverse, a searchable electronic database that contains manufacturer data of over 4000 different materials [25]. This experimental dataset covers 14 broad categories: physical, application composition, mechanical, impact, processing, bio-data, absorption, geo-economic, thermal, electrical, optical, durability, and environmental measurements. This broad and comprehensive collection of commercial data makes the CES Selector a powerful tool for analyzing a wide range of materials. CES Selector was chosen for its wide and comprehensive evaluation of experimentally collected measurements spanning across various material classes. The CES Selector selectively contains information from commercial materials provided by manufacturers and suppliers. Due to this focus on commercialized materials, results from SOM analysis on the CES Selector dataset will reveal information about potential materials of interest that are already commercially applicable. This makes CES Selector very valuable as a tool for establishing relationships between materials that are important in commercial and industrial uses.

The CES Selector provides a data extraction tool that allows the user to select individual materials for property value comparisons. This tool allows for the selection of many different materials into a *.csv* format that can be saved onto a computer for quick processing.

3.2. Data processing and cleaning

Among various types of property data available in CES Selector, we focused on only mechanical, thermal, electrical, and optical properties in this paper. Other properties were not used in our analysis, as they are mostly categorical and do not have numerical values that can be extracted by our SOM method. MaterialUniverse contains several categories of materials such as metal, ceramic and glass. To narrow down the dataset for computational analysis, only two classes of materials were analyzed: ceramics and metals.

Materials for the training dataset were selected randomly within the two classes of our interest (ceramics and metals). By using the data extraction tool in Granta, a total of 421 materials with their various properties were produced. The training data consists of 196 ceramic materials and 225 metallic materials. Special attention was given to avoid bias in data caused by reporting materials with very similar alloy compositions, which is a primary concern for metals. To avoid this bias, metals were chosen with different base materials, different dopants, and different processing methods to diversify the training dataset.

The raw data extracted from Granta, however, contains some missing values. Materials containing too many missing values and/or properties were carefully removed without causing bias. This process eliminated all the "holes" in the training data set, which means data padding is not necessary. In this way, the cleaned training dataset now contains 398 materials and 59 properties. Some of the properties are numerical data, e.g. Young's modulus, and some of the properties are categorical data, e.g. magnetic or non-magnetic. Only the numerical properties can be used to train the SOM. Some properties were also removed (such as toughness and thermal shock resistance) because they were derived from calculations on data already contained in Granta. Thus, 398 materials and 21 numerical properties were input into SOM as training data. Though the categorical properties were not used to train the SOM, it still can be visualized in our clustering maps and play an important role in the analysis of the clustering result. Cluster maps are produced using an algorithm described in Section 3.3 and depict materials with similar property values grouped together.

Granta reports all 21 of the numerical property data of interest as a range of values. Our SOM algorithm requires that only single values are used, rather than a range of values. For these cases, we tried 2 methods to process the data. In the first method, we took the arithmetic average of the range, which was calculated by dividing the sum of the minimum and maximum value in the range by 2. This arithmetic average replaced the range reported as the value to be used for our SOM analysis. A concern to this method is that the average is not representative of the material, especially for properties with wide ranges. In an effort to normalize the range-recorded values to be more representative of the dataset, the square root of the product of the minimum value and the maximum value in the range was used as the second method. All of the training data were processed similarly for both cases. It was found that SOM analysis on both methods (arithmetic average vs. square root average) yielded no significant differences between the so-called "cluster maps" that were generated from our training data (see Fig. 3). Thus, we chose to use the arithmetic average method to preprocess the training data in our later work for simplicity. Although cluster maps will be analyzed further in section 4 of this paper, it is important to recognize at this stage that the geometry of the clusters (shown as the boundaries in Fig. 3) are consistent and the position of the materials are only slightly changed. Therefore, we conclude that the clustering results are not sensitive to what averaging method we choose to use.

These cluster maps that are produced by our SOM method is not only a powerful visualization tool, but also has an advantage in that it can be used to represent both numerical as well as categorical data (as shown by our use of color-coordinated dots to represent different material classes in section 4 of this paper).

3.3. Training SOM and visualizations

All the algorithms included in this paper were realized in Python 3.6 environment. The training and visualization of SOM was implemented by a python package called Tfprop_sompy. It was developed by Kikugawa and Nishimura, based on an open-source SOM package called SOMPY [26]. The data cleaning, preprocessing and visualization work also involve some commonly-used python packages including Pandas, Numpy, Scipy, Scikit-learn, Seaborn and Matplotlib.

While training a SOM, the following parameters should be specified for the python function: normalization method of training data, map size.

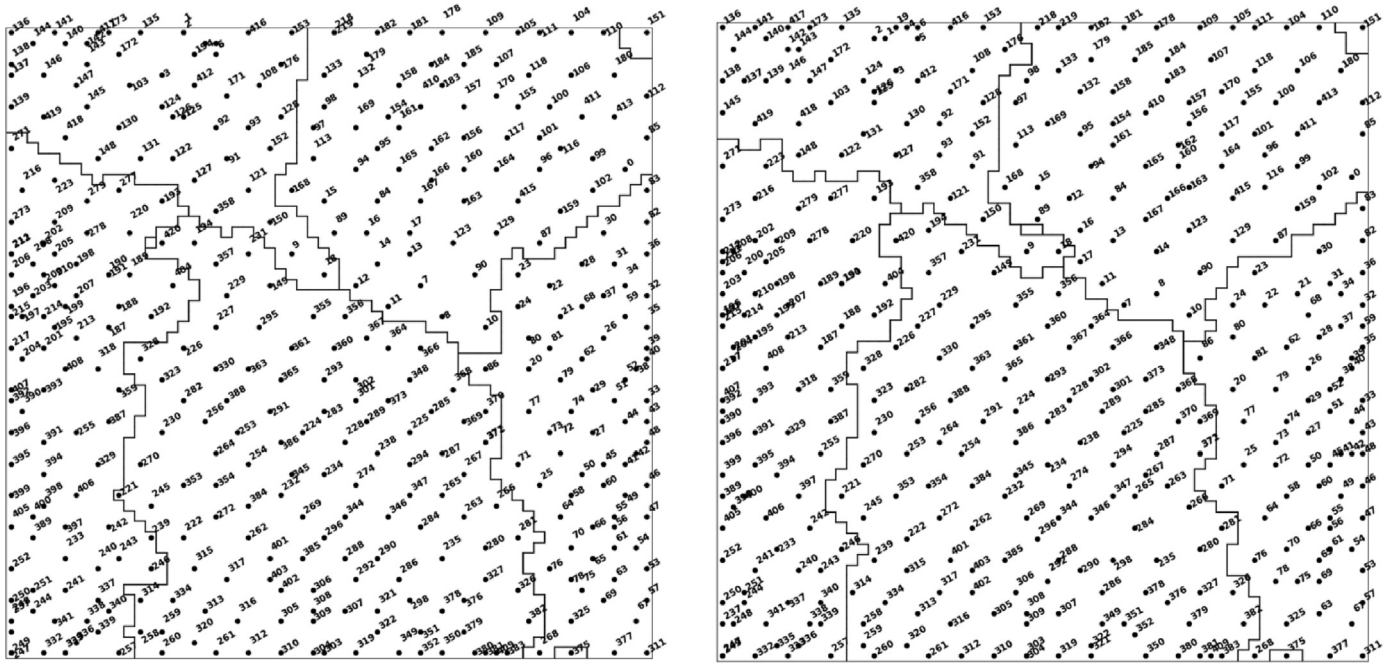

Fig. 3. Cluster maps produced by SOM algorithm on training dataset containing 398 materials and 21 properties. Training data presented in range values were calculated by taking the average of the minimum and maximum values in the range (left) and by taking the square root of the product of the minimum and maximum values in the range (right). Each dot and corresponding number index refer to an individual material that is used in our training data.

In our work, we standardized each of the training data by $x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, in which x_{ij} represents the j-th property of the i-th material in our training dataset, μ_j, σ_j represent the mean and standard deviation of j-th property of all the materials. An appropriate map size is expected to output a SOM where each node in it matches with no more than one training data point. The map size is adjusted based on this rule. For all our later work, we used a map size of 60 × 60.

The training process of SOM has two stages: firstly roughtune then finetune. In the roughtune stage, the neighborhood radius gradually decreases from 8 to 2, and this stage lasts for 273 times of iteration. In finetune stage, the neighborhood radius gradually decreases from 2 to 1, and this stage lasts for 361 times of iteration.

When running on a local computer with a 2.3 GHz two-core processor, the training process of our SOM with a map size of 60 × 60 on our dataset with 398 materials and 21 properties takes 221.513 s. For a different project in our lab, when running our algorithm on a local computer with a dataset of 739 materials with 5 features, the training process takes 57.675 s. This computer has one processor with a 2.00 GHz processing speed.

In our work, we employed U-matrix map, clustering map and heatmaps to visualize the SOM we trained [27]. The U-matrix map was produced by mapping the average Euclidean distance between each node and its nearest neighbors in high-dimensional space onto the two-dimensional map, as shown in Fig. 4. For example, the value of node 5 on the U-matrix map is given by the average of the Euclidean distance between node 5 and its nearest neighbors: node a, b, c, d. The values on the U-matrix map reflects the 'similarity' of each nodes and their neighbors. The high U-matrix value area on the U-matrix map, which are shown by red color region in Fig. 7(a), means the nodes as well as the training data points in this area are far away from each other in the high-dimensional space, and in other words, they are very 'different'. For training data points in the blue region, they are very 'similar' to each other.

To produce a clustering map, we first performed the K-means clustering on the SOM we trained, as illustrated in Fig. 5 [28]. K-means clustering is a simple way of partitioning all the nodes in SOM into K distinct, non-overlapping clusters. It starts with randomly setting the

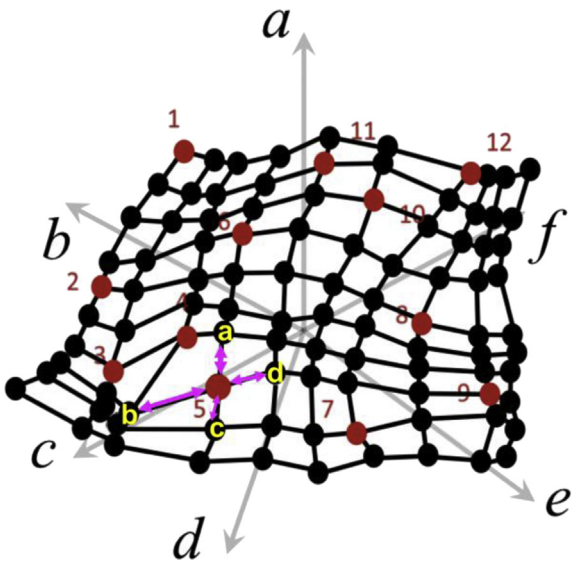

Fig. 4. The average distance in high dimensional space determines the value of U-matrix.

initial K centroids for the K clusters, as shown in Fig. 5(a). Then iterate the following two steps until the clusters assignments stop changing: 1) assign each node to the cluster whose centroid is closest (shown in Fig. 5(b)), 2) calculate the new centroids by taking the mean value of all of the samples assigned to each previous clusters (shown in Fig. 5(c)). Finally, the clustering result can be visualized with the SOM, as shown in Fig. 5(d). The number of clusters are empirically determined and then validated by the U-matrix map we got. A reasonable clustering map is expected to have cluster boundaries overlapping with red regions on the U-matrix map (Fig. 7(a)).

Heatmaps are then produced by separately mapping each element of the node's weights onto the two-dimensional map (notice that node's

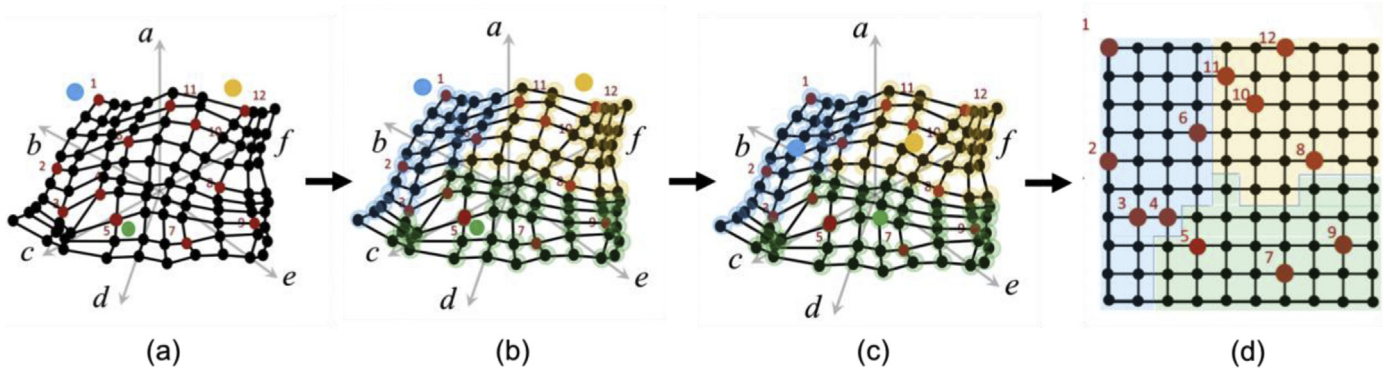

Fig. 5. Perform K-means clustering method on the trained SOM.

weights are de-standardized before producing the heatmaps). For example, as illustrated in Fig. 6 below, by mapping the 'a' axis element of node's weights, we are able to obtain distribution of the *a*'s value on the two-dimensional grid, and in the same way we can get the distribution of b to f's value. The color bar of the heatmap was set in a way that the red color represents the highest value on the map and the blue color represents the lowest value. The advantage of the method is that the variation of value on the map is most significant, which can help the users detect the correlation between properties.

It seems that the total number of heatmaps we can get equals to the number of elements that the node's weight contains (i.e. the number of the variables that the training data has). However, since a heatmap is essentially a matrix, we can produce a new heatmap by doing the element-wise operation on two or even more heatmaps we already have. For example, if we want to know the distribution of $(a \bullet b)$, we can do element-wise multiplication on the heatmaps *a* and *b* in Fig. 6. This method is useful in computing new data not originally provided in our training dataset.

4. Interpretation of SOM

4.1. Clustering map

As mentioned in section 3.2, we studied material properties such as mechanical properties like Young's modulus, Poisson ratio, and fracture toughness, thermal properties like thermal conductivity, and electrical properties like electrical conductivity for 398 different commercial metal and ceramic materials. The U-matrix map and clustering map are shown in Fig. 7a and b. The cluster map (Fig. 7(b)) indicates the different clusters of similar materials after running the SOM algorithm on our dataset. The boundaries in the cluster map that is produced closely resembles the red region in the U-Matrix (Fig. 7(a)), validating the

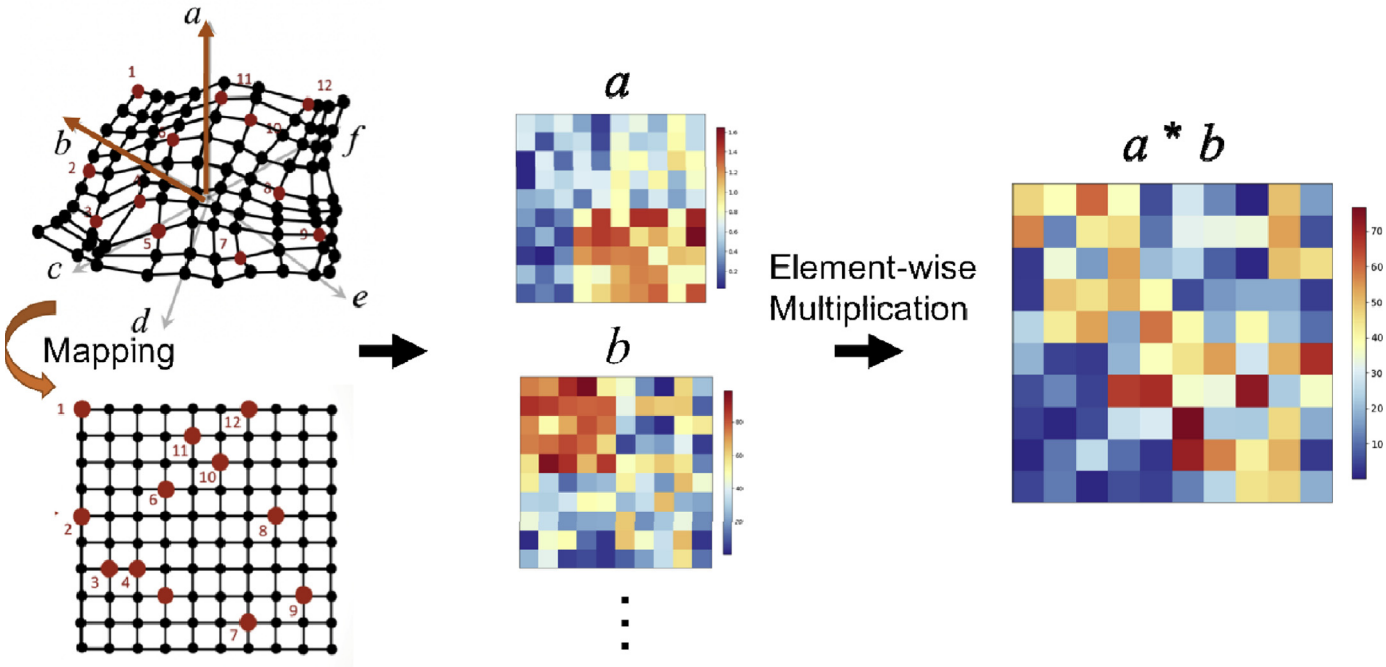

Fig. 6. Produce heatmap by mapping each element of the node's weights onto the two-dimensional map.

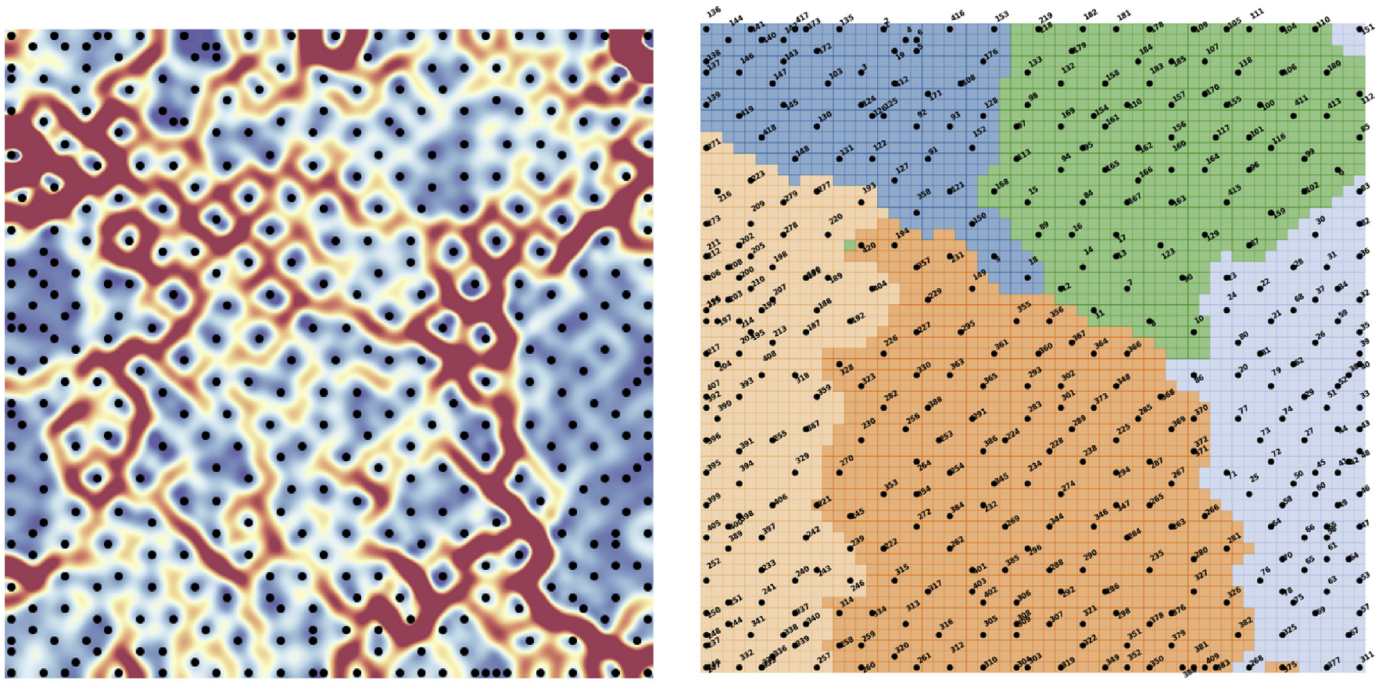

Fig. 7. a. U-Matrix produced from SOM algorithm indicating boundaries between material clusters. b. Cluster map produced from SOM algorithm containing 398 materials represented by individual dots, using averaged values.

clustering result. The cluster map contains five distinct clusters represented by different colors, and the dots on the cluster map mostly represents an individual material (a few dots represent two materials). In addition, the materials we studied belong to different material categories such as ceramic and metal. The numbers labeled on the cluster maps are ID numbers that each correlate to a unique material in our training dataset.

During our data extraction from Granta, there were many categorical data that we could not apply to our SOM algorithm because they are not numerical values (therefore, we removed these properties in our training dataset from 59 to 21 properties as explained in section 3). Therefore, to add an extra layer of interpretability to our SOM analysis, we labeled each dot by a different color with categorical information from Granta such as magnetic/non-magnetic, resistance of a material to degradation by strong acid, material class and so on (shown in Fig. 8).

This utility of SOM is powerful because it allows us to not only visualize numerical data on the maps, but also non-numerical data. This capability adds another layer of analysis to our SOM interpretability and help us identify how materials in the same category are similar or dissimilar. There are several categorical data from Granta that we had excluded from our data processing step because they are non-numerical. However, by using this method of introducing non-numerical data to our cluster maps, we can expand upon the wide range of property information provided by Granta and make deeper interpretations to our maps.

There are several interesting observations in analyzing the cluster maps produced by labeling categorical data in Fig. 8. It appears that the magnetic ability (Fig. 8(a)) and resistance of materials to strong acids (Fig. 8(b)) show no clear correlation between materials because the colored dots do not display a noticeable trend in their clustering location. However, the cluster map representing materials by their material class (Fig. 8(c)) does provide some interesting material relationships. It is demonstrated in this cluster map that materials in similar classes are usually located within the same regions, as is indicated by similar dot colors being found in similar areas. For example, non-ferrous metals are found in the bottom-left corner and technical ceramics are found in the upper-right corners. Ceramics are very clearly separate from metals and non-metals, which supports our understanding of the differences between the fundamental bond structure of these materials. Therefore, this method of representing categorical data using our SOM method can reveal material relationships that are not represented numerically.

Fig. 9 is an extension of Fig. 8, in which we layer two types of categorical data in our SOM maps. As in Fig. 8, the dots are color-coded to represent their material class however additional labels next to each data point were added to indicate their base material. It is shown that materials with similar base material tend to be located within the same region of the cluster map. This is clearly seen in oxides being represented in the upper-right corner and carbon-based materials being represented in the upper-left corner. Similar base materials located in the same region in our cluster maps validate that results produced by our SOM code are consistent with our expectations that the base material provides a foundation for the overall properties of the material. Upon further visual analysis of the location of similar materials in the cluster maps, several interesting materials stand out that are located farther away from other similar materials. This SOM method allows us to quickly visualize correlations between material structure-properties in an intuitive way that is very supportive of students' educational training in identifying similar or dissimilar materials.

Therefore, we have identified and labeled several interesting materials that are outliers compared to other similar materials in their class in Fig. 9. Several materials of interest are circled. The two beryllium materials circled in red are unique because they are found in the cluster that is occupied by ceramic oxides and carbides, while other non-ferrous metals (indicated by the light green points) are found in different regions. The identity of these two beryllium metals are Beryllium, grade 0–50 that is hot isostatically pressed and Beryllium, grade I–220B that is vacuum hot-pressed. These materials are used mainly for aerospace engineering as structural material for projectile and optical components. Our SOM method provides a visual verification of how unique processing

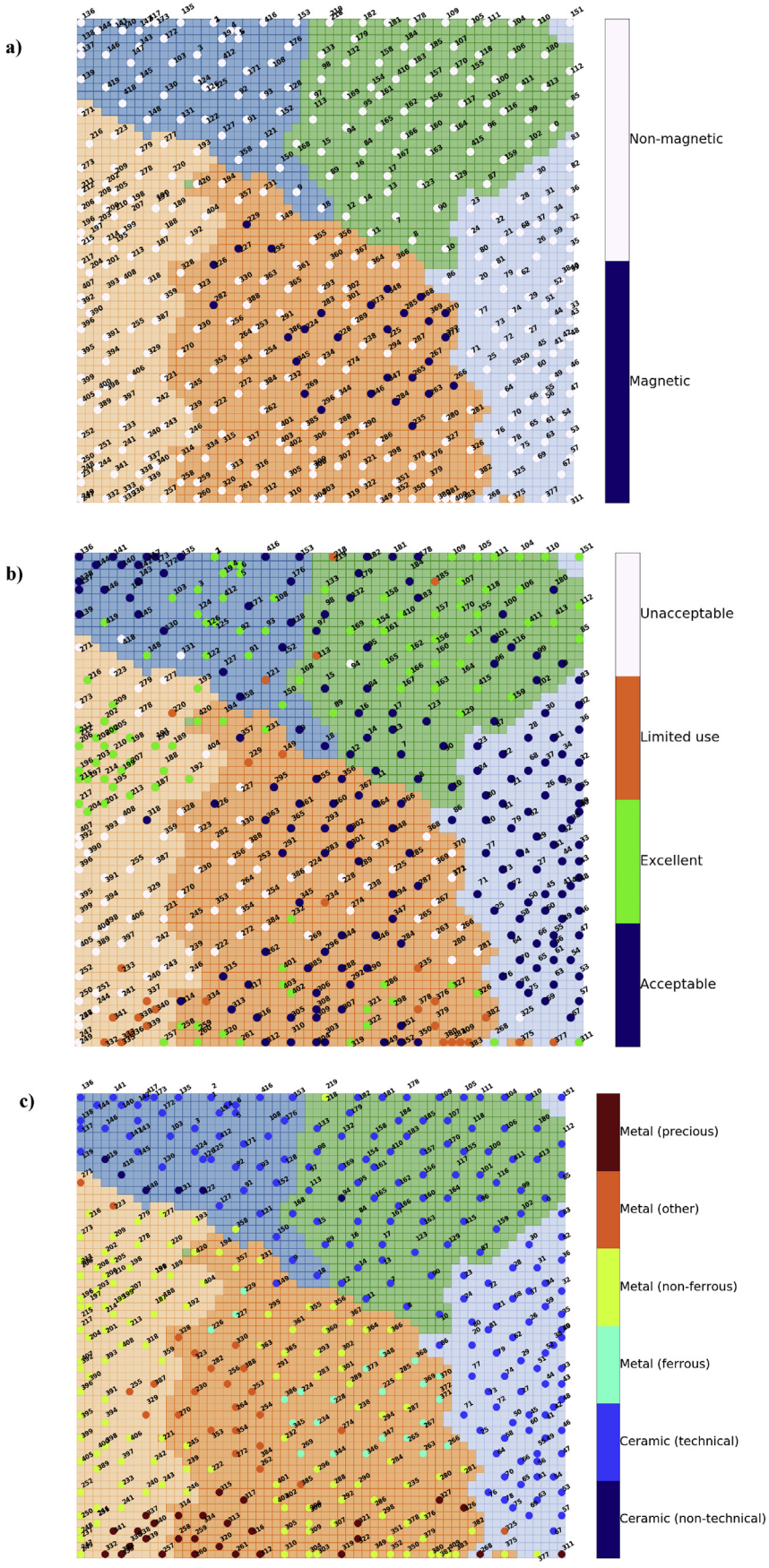

Fig. 8. Cluster map produced from SOM algorithm containing 398 materials. The materials are individually represented by points that are color-coordinated to indicate the magnetic ability (a), the tolerance to strong acid (b), and material class (c). The cluster map representing magnetic ability portrays each material as either non-magnetic (white) or magnetic (dark blue). The cluster map representing resistance to degradation by strong acids portrays each material as intolerant (white), limited tolerance (orange), excellent tolerance (green), or acceptable tolerance (dark blue). The cluster map representing different material classes portrays each material as a precious metal (red), non-ferrous metal (green), ferrous metal (teal), other metal (orange), technical ceramic (light blue), and non-technical ceramic (dark blue). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

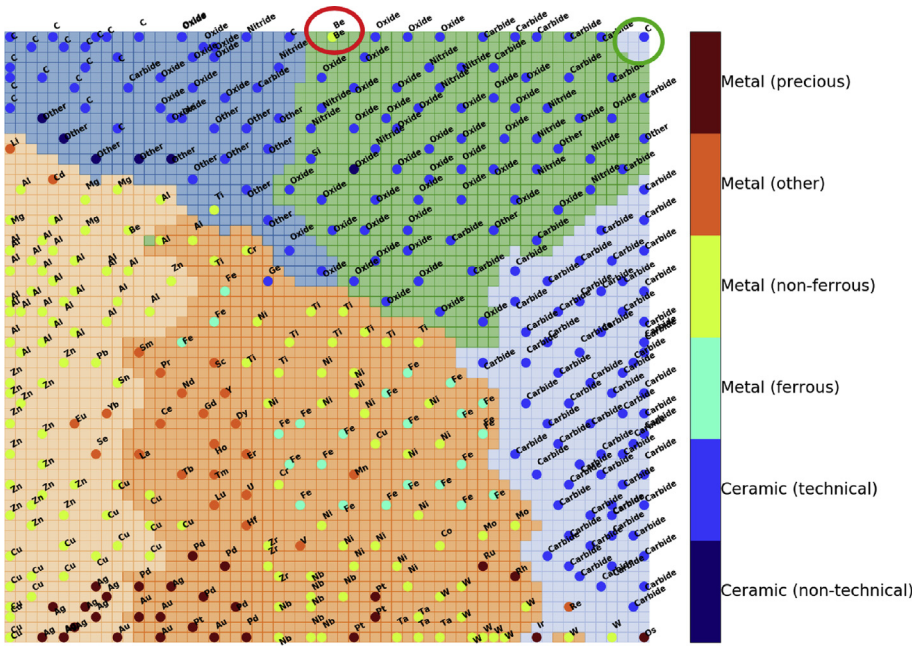

Fig. 9. This cluster map is an extension of Fig. 8, where we not only color-labeled the dots with material class type, but the base material is indicated as well. Several 'outlier' materials are identified and circled in different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

methods can result in materials that have vastly different properties than other materials in the same class. In the case of the two beryllium materials, they have properties more similar to ceramics than other metals as a result of being hot-pressed.

The green circled material in the upper right corner of Fig. 9 is a ceramic with a carbon base. This material is in a separate cluster away from other carbon-based materials and is different from the surrounding materials. This can be verified by the U-matrix map in Fig. 7(a), where the material is surrounded by red boundary. The identity of this material is diamond. It is important to note that although this point is located in a separate cluster from the other surrounding points, it actually belongs to the same light blue cluster as the other carbide-based materials in the lower right in Fig. 9.

4.2. Heat maps

As described in section 3.3, heatmaps are produced using SOM by separately mapping each element of the node's weights onto the two-dimensional map. Shown in Fig. 10 are 21 heat maps generated from the 21 properties from our training data, from which valuable materials information can be gleaned from interpreting relationships between different material properties. For example, properties with similar heat maps (Young's modulus, Bulk modulus, Shear modulus, Compressive strength, Hardness, and Flexural modulus) are highly correlated. Since these properties are already known to be highly correlated, we can see that SOM is a powerful and effective tool in validating the expected correlation between different material properties and to predict unknown correlations. The advantages of using SOM to generate these heat maps is for quick, visual, and qualitative understanding of relationships between material properties. Material properties with high correlation will have similar heat maps, with similar red and blue regions.

A correlation matrix was used to quantify dependencies between two material properties. For this, a Pearson correlation matrix was calculated for all 21 properties that were analyzed in the training data [34]. The correlation value between two properties is calculated using the equation:

$$\frac{r_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\ \ \bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} \tag{2}$$

where r_{xy} is the Pearson correlation coefficient between property X and property Y, x_i and y_i are the value of property X and property Y of the i-th material ($i = 1, 2, \ldots, n$), n is the total number of materials in the training dataset, $\bar{x}$ and $\bar{y}$ are the mean value of property X and property Y of all the materials. A correlation value between two properties that has an absolute value of greater than 0.7 is considered as 'high' and therefore is highly related to each other. The purpose of this analysis is to establish that two properties are correlated to each other, not that they are necessarily causal. The sign of the actual value (+or -) provides information about whether two properties are positively or inversely related to each other. For example, the correlation value between flexural modulus and Young's modulus is 0.99, meaning that there is a direct positive relationship between these properties. This is logical because the flexural modulus of a material measures the resistance of a material to resist bending, and materials with higher Young's modulus are by definition more resistant to deformation under load. Alternatively, a negative correlation value, such as that between thermal expansion coefficient and melting temperature (−0.70) represents properties that are strongly inversely related. This relationship is described by understanding fundamental knowledge about bonding strength. The stronger the bond strength in a material, the lower the thermal expansion coefficient and the higher the melting point. Both cases of positive and inverse relationship between properties are shown in Fig. 11. The heat maps allow for quick identification of correlation relationships based on visual analysis of the locations of the red and blue regions between heat maps. The correlation matrix and heat maps generated across the 21 properties support expected material trends, and is a valuable tool in materials education as a visual representation of these trends.

While the correlation matrix and the heat maps obtain the same conclusion, the heat maps are able to provide further information about the distribution and localization of correlated properties. Our method of generating heat maps is capable of visualizing the correlations between multiple properties, providing a broader analysis than using the

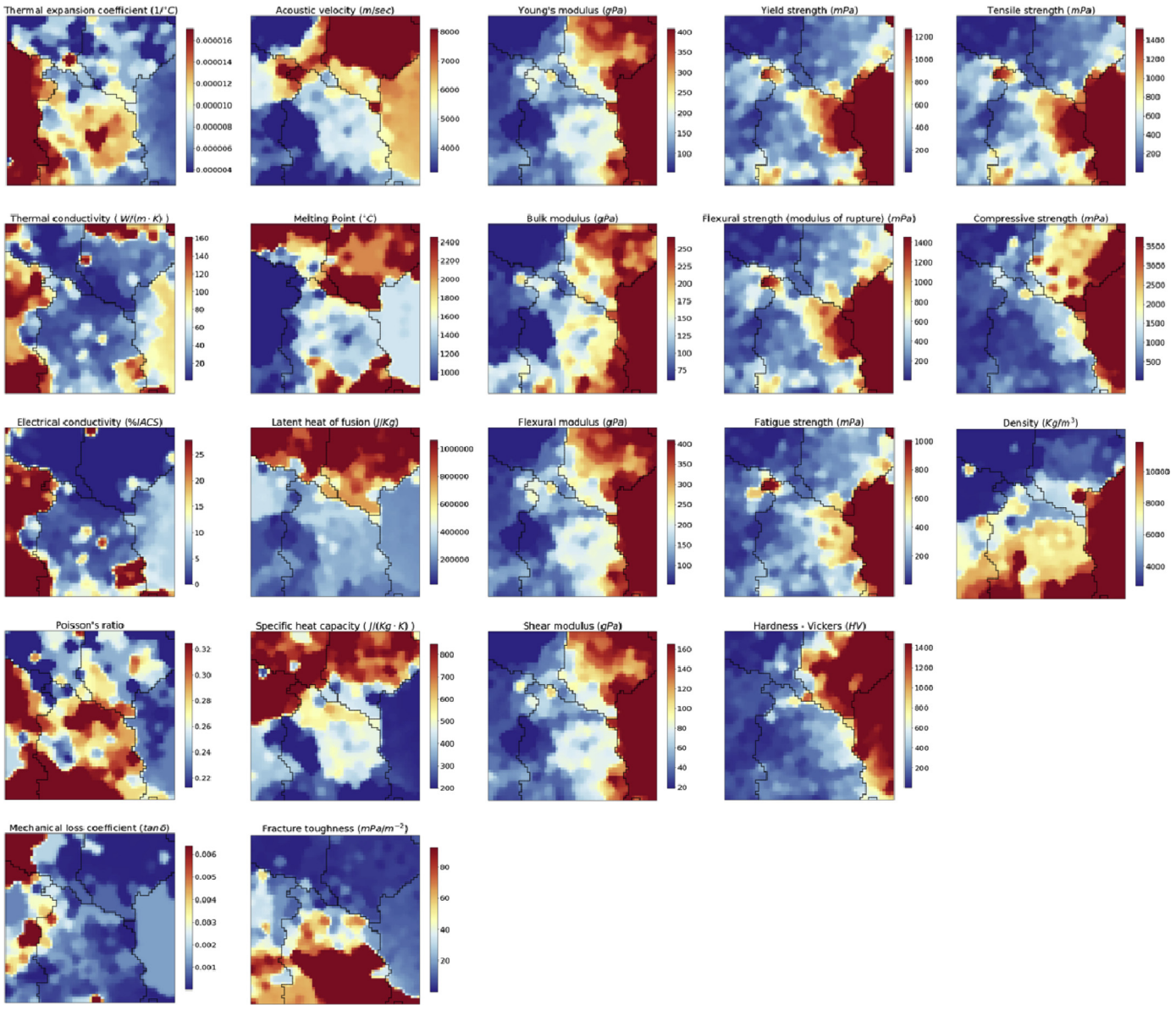

Fig. 10. Heat maps produced from SOM code that are used to visualize similarities among various properties of the 398 materials evaluated. There are 21 heat maps to represent the 21 different material properties that were in the training dataset. The red regions indicate high values and blue regions indicate low values. The heat maps are arranged such that correlated properties are grouped together in columns. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

correlation function alone.

However, the heatmaps for flexural modulus and Young's modulus (right) show a positive relationship between these two material properties. The regions with high flexural modulus (red) also correspond to high values in the Young's modulus heat map.

The heat maps are in accordance with known material equations as demonstrated in Fig. 12, which reinforce existing knowledge about mechanical properties. We compared the relationship between the bulk modulus, Young's modulus, and Poisson ratio. These three properties are related through Equation (3):

$$K = \frac{E}{3(1-2\nu)} \tag{3}$$

where K is the bulk modulus, E is the Young's modulus, and ν is the Poisson ratio. As shown in Equation (3), the bulk modulus and Young's modulus is positively correlated. The Poisson's ratio is relatively consistent among materials, so it has little impact on the positive relationship between the bulk and Young's modulus. In this way, SOM validates visual results and correlations with mathematical relationships between material properties. This type of mathematical analysis supplements SOM as an educational tool by verifying correlational properties between materials with a visual output. However, SOM is not restricted on only correlated properties; it can be used on very uncorrelated properties as well. Other applications for the heat maps will be shown in the next sections in this paper, showing the usefulness of our SOM method for different applications. In this way, SOM validates visual results and correlations with mathematical relationships between material properties and is a strong educational tool.

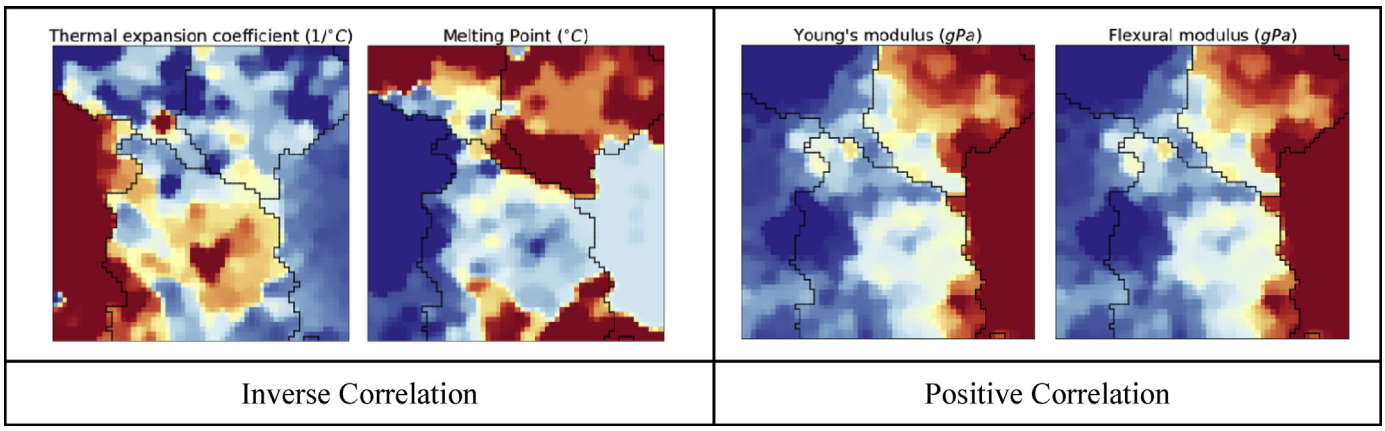

Fig. 11. The heatmaps for thermal expansion coefficient and melting point (left) show inverse relationship between these two material properties. For these properties the regions with high thermal expansion coefficient (red) correspond to low values (blue) in the melting point heat map. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

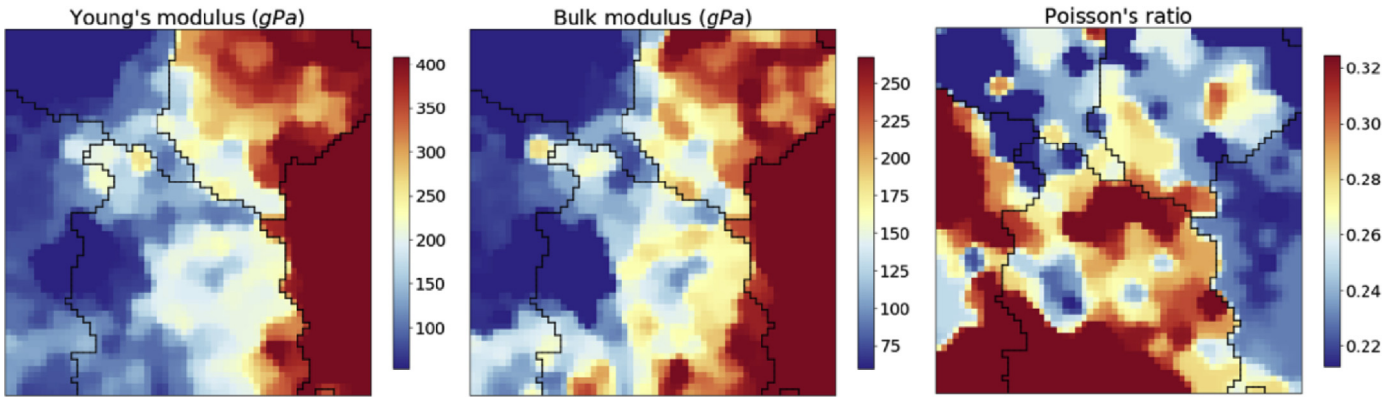

Fig. 12. Heat maps of Young's modulus (left), bulk modulus (center), and Poisson ratio (right). These maps show that the bulk modulus and Young's modulus are positively correlated because the heat maps look similar. On the other hand, these maps also show that the Poisson's ratio is relatively consistent among materials, so it has little impact on the positive relationship between the bulk and Young's modulus.

5. Applications of the SOM method

5.1. Unique materials in terms of thermal and electrical conductivities

Using a similar method to the previous example comparing red and blue regions of the heat maps, we identified materials with unique material property relationships (Fig. 13). We have labeled three groups of unique ceramic and metal materials and show these materials with different marker and color on the heat map. Most ceramics have low thermal and electrical conductivity. However, in group 1 (magenta triangles) these ceramics have high thermal conductivity and low electrical conductivity. The identities of these materials are silicon carbide, silicon, graphite, diamond, beryllia, and aluminum nitride. Some of these materials, such as silicon carbide and aluminum nitride are typically known as 'fine ceramics,' which are different from conventional ceramics because they are engineered to have precise control of composition [30].

In group 2, green triangles indicate that these materials are the only two ceramics in our training data that have high thermal and electrical conductivity. Thermal and electrical conductivity are related by the Wiedemann-Franz law. This law which states that the relationship between thermal and conductivity is related to temperature and is based on the fact that both heat and electron transport both involve the movement of free electrons [35]. One of the materials is graphite (parallel to plane), which is the only ceramic in our training data that has both high thermal and electrical conductivity. It should be noted that our training data contains measurements of graphite (pyrolytic) collected from two different orientations: parallel to plane (ID 136) and perpendicular to plane (ID 135). We treated the different orientations as two data points. It is known that when graphite is measured perpendicular to the plane, it has much lower thermal and electrical conductivity compared to when it is measured parallel to the plane [32]. There are two more materials of interest (ID 218 and ID 219) in group 2. These materials are both Beryllium alloys, and we have discussed the uniqueness of these two metals in Section 4.1.

Most metals in our training data have a positive correlation between thermal and electrical conductivity. For group 3 (cyan triangles), these metals are unique because they have high electrical conductivity and low thermal conductivity. The identities of these two aluminum-based materials are: Al–47%SiC(f) and Al–50%B(f). These two metals are compositionally unique because they only contain about 50% aluminum. The high level of dopants in the aluminum metals give these materials unique properties compared to other aluminum metals in the training data. Al–47%SiC(f) (commercial name: Al 6061-SCS-2) is used mainly for aerospace and bridge components, while Al–50%B(f) (commercial name: Al 6061-B) is used mainly for aerospace components.

From this example, we demonstrate that SOM can be used to identify unique materials that disobey the positive correlation between thermal conductivity and electrical conductivity that is true for most materials.

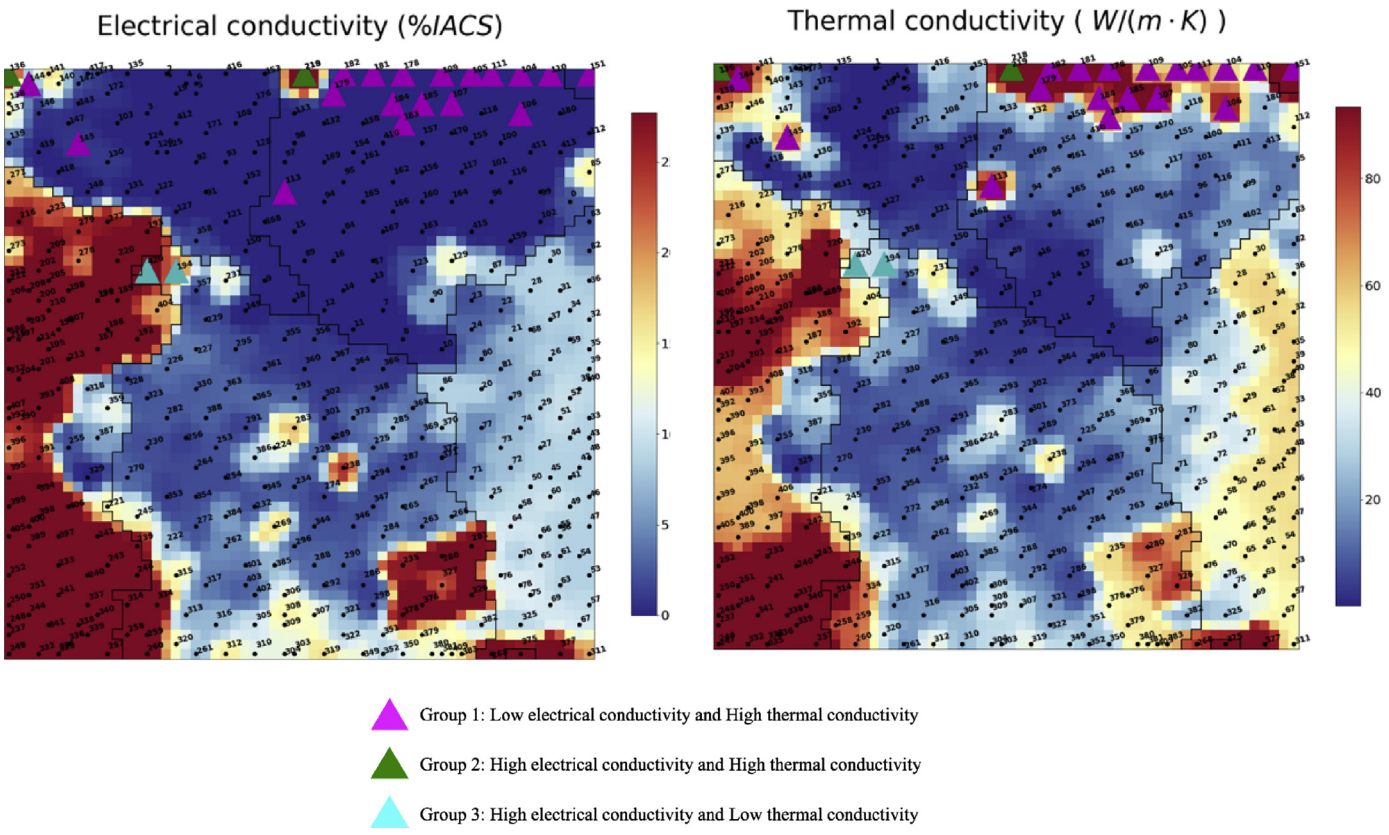

Fig. 13. Heat maps of thermal and electrical conductivity generated by SOM algorithm on 398 materials and 21 material properties. The colored dots indicate materials of interest that have unique material relationships.

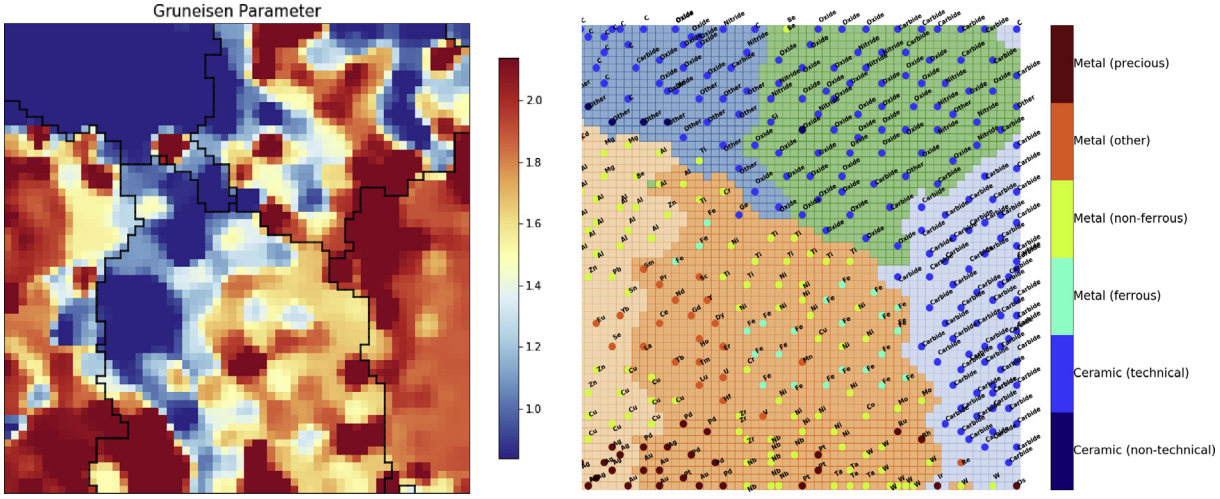

Fig. 14. Heat map of Grüneisen parameter generated by element-wise multiplication (left), the clustering map of all training data (right).

5.1.1. Evaluation of the Grüneisen Parameter

In addition to visualizing correlations between material properties, our SOM method can be used to evaluate new data that is not part of the original dataset. Since the SOM heat maps are essentially matrices, it is shown that they can be combined using element-wise multiplication to produce new heat maps representing different properties not part of the original dataset. Here is an example of using this method of element-wise multiplication to produce a heat map for the Grüneisen parameter (Fig. 14). The Grüneisen parameter measures the magnitude of the dimensional change in a system as a response to the thermal energy variation induced by changing temperature. This parameter connects two important physical properties of a solid: the specific heat capacity (C_p)

and the volume thermal expansion coefficient (α). It is calculated using the equation:

$$\gamma = \frac{\alpha K}{\rho C_\rho} \tag{4}$$

where γ is the Grüneisen parameter, α is the volume thermal expansion coefficient, K is the bulk modulus, ρ is the density, and C_p is the specific heat capacity.

To visualize the map for the Grüneisen parameter, the heat maps for thermal expansion coefficient (α), Young's modulus, density, and specific heat are combined through element-wise multiplication to produce Fig. 14. The Grüneisen Parameter on our heat map varies from 0.8 to 2.2, so the value range is relatively small, which agrees with current study of Volume-Based Thermodynamics and Thermoelasticity (VBT) [33]. The Gruneisen Parameters for each material in our dataset were calculated and represented on the heatmaps, which were color coordinated to represent the distribution of value range. It is shown that the materials in upper left cluster and lower right cluster have relatively constant Grüneisen Parameter. Comparing the clustering map to the heat map for the Grüneisen parameter, it is shown that precious metals (dark red dots) have high Grüneisen parameters, while the orange dots characterizing 'other metals' have low Grüneisen parameters.

6. Conclusion

In our work, a SOM has been trained on a high-dimensional dataset containing 21 numerical properties of metals and ceramics to evaluate materials' property-property relationships. Although we only evaluated 398 materials, we have shown incredible relationships between material properties, such as mechanical properties. We have also shown multiple layers of interpretation in our SOM method through visualizing not only numerical properties but also categorical information of materials. This supports SOM as a powerful tool for materials research and education, even with small datasets. Our analysis of the SOM method expanded over several applications.

Cluster maps generated using SOM validated the similarities between materials in the same material class as well as materials containing the same base material. The cluster maps revealed several 'outlier' materials that were categorized in different clusters compared to similar materials. We have identified these 'outlier' materials to be: two Beryllium metals and diamond. These materials are unique due to their processing method and base material composition, supporting SOM as an effective tool for identifying material similarities and dissimilarities. We demonstrated that SOM can be applied to generate heat maps that can be paired with Pearson correlation coefficients to visualize both positive and inverse correlations between material properties. For example, we showed that flexural modulus has a positive correlation to Young's modulus while the thermal expansion coefficient and melting temperatures were inversely correlated due to bonding structure. The heat maps also supported the mathematical relationship between bulk modulus, Young's modulus, and Poisson ratio. Furthermore, we compared heat maps produced by SOM to quickly identify 3 groups of unique materials that disobey the positive correlation between thermal conductivity and electrical conductivity. The first group are the ceramics with high thermal conductivity but low electrical conductivity, which contains silicon carbide, silicon, graphite, diamond, beryllia, and aluminum nitride. The second group are the ceramics with both high thermal and electrical conductivity. One of them is graphite, and the others are Beryllium alloys which we already identified in the analysis of cluster maps. The third group are the metals with low thermal conductivity but high electrical conductivity, which are two aluminum-based materials.

We also demonstrated that heat maps can be manipulated through element-wise multiplication to evaluate the value of Grüneisen parameter, which is a quantity not included in the training data. This technique provides a deeper analysis of our training data by visualizing computed values not included in our original dataset.

In addition to conventional Ashby maps, we illustrated that SOM is another visualization technique for materials science education and research while it emphasizes different aspects of the data. Ashby maps allows for showing the ratio of the two properties of materials, while SOM allows for the quick and intuitive visualization of many properties of materials at the same time. Together, the clustering and heat maps produced from SOM can be used to support educational experiences through visual analysis of material properties, and validate existing knowledge about material properties relationships. In conclusion, SOM is an integrated approach to study material property relationships and identify materials with uniqueness.

Acknowledgements

The authors would like to acknowledge encouragement and support made through the University of Washington-Tohoku University: Academic Open Space (UW-TU:AOS). The authors also appreciate Cross-Ministerial Strategic Innovation Promotion (SIP) Program of Japan for providing the funding of purchasing Granta database.

References

[1] T. Kalil, C. Wadia, Executive Office of the President, National Science and Technology Council, *Materials Genome Initiative for Global Competitiveness*, 2011.
[2] D.A.C. Beck, J.M. Carothers, V.R. Subramanian, J. Pfaendtner, Data science: accelerating innovation and discovery in chemical engineering, AIChE J. 62 (5) (2016) 1402–1416, https://doi.org/10.1002/aic.15192.
[3] https://www.nist.gov/about-nist.
[4] Workshop on Materials Science and Materials Engineering Education: National Science Foundation, the Future of Materials Science and Materials Engineering Education, National Science Foundation, Arlington, 2008.
[5] K. Rajan, Materials informatics, Mater. Today 8 (10) (2005) 38–45, https://doi.org/10.1016/S1369-7021(05)71123-8.
[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.
[7] T. Bäck, J.N. Kok, G. Rozenberg, Handbook of Natural Computing, Springer, Berlin, 2012.
[8] V.G. Maltarollo, K. Honório, A. Ferreira da Silva, "Applications of Artificial Neural Networks in Chemical Problems," Artificial Neural Networks-Architectures and Applications, IntechOpen, 2013, https://doi.org/10.5772/51275.
[9] Richard D. Lawrence, George S. Almasi, Holly E. Rushmeier, A scalable parallel algorithm for self-organizing maps with applications to sparse data mining problems, Data Min. Knowl. Discov. 3 (1999) 171–195, 2.
[10] I. Jolliffe, Principal Component Analysis, Springer, Berlin Heidelberg, 2011.
[11] M. Kasslin, J. Kangas, J. Simula, Process state monitoring using self-organizing maps, Artif. Neural Netw. (1992) 1531–1534, https://doi.org/10.1016/B978-0-444-89488-5.50152-4.
[12] J. Lampinen, E. Oja, Distortion tolerant pattern recognition based on self-organizing feature extraction, IEEE Trans. Neural Netw. 6 (1995) 539–547, https://doi.org/10.1109/72.377961, 3.
[13] J. Heikkonen, P. Koikkalainen, E. Oja, Self-organizing maps for collision-free Navigation, Proc. World Congr. Neur. Networks WCNN 3 (1993) 141–144.
[14] T. Kohonen, Combining linear equalization and self-organizing adaptation in dynamic discrete-signal detection, in: IJCNN International Joint Conference on Neural Networks, IEEE, 1990, https://doi.org/10.1109/IJCNN.1990.137573.
[15] T. Kohonen, Engineering applications of the self-organizing map, Proc. IEEE 84 (10) (1996) 1358–1384, https://doi.org/10.1109/5.537105.
[16] Suwardi Annas, Takenori Kanai, Shuhei Koyama, Principal component analysis and self-organizing map for visualizing and classifying fire risks in forest regions, Agric. Inf. Res. 16 (2007) 44–51, 2.
[17] M. Ashby, Mater. Sel. Mech. Des. 3 (1999).
[18] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, New York, 2017.
[19] Y. Bengio, Learning deep architectures for AI, Found. Trends Machine Learn. 2 (1) (2009) 1–127, https://doi.org/10.1561/2200000006.
[20] G. Kikugawa, Y. Nishimura, K. Shimoyama, T. Ohara, T. Okabe, F.S. Ohuchi, Data analysis of multi-dimensional thermophysical properties of liquid substances based on clustering approach of machine learning, Chem. Phys. Lett. 728 (2019) 109–114, https://doi.org/10.1016/j.cplett.2019.04.075.
[21] Y. Oya, G. Kikugawa, T. Okabe, Clustering approach for multidisciplinary optimum design of cross-linked polymer, Macromol. Theory Simul. 26 (2) (2017), https://doi.org/10.1002/mats.201600072.
[22] R. Liu, HDAT: web-based high-throughput screening data analysis tools, Comput. Sci. Discov. 6 (1) (2013), 014006, https://doi.org/10.1088/1749-4699/6/1/014006.
[23] P. Schneider, A.T. Müller, G. Gabernet, A.L. Button, G. Posselt, S. Wessler, J.A. Hiss, G. Schneider, Hybrid network model for "deep learning" of chemical data:

application to antimicrobial peptides, Mol. Inf. 36 (1–2) (2016) 1600011, https://doi.org/10.1002/minf.201600011.
[24] R. Jha, G.S. Dulikravich, N. Chakraborti, M. Fan, J. Schwartz, C.C. Koch, M.J. Colaco, C. Poloni, I.N. Egorov, Mater. Manuf. Process. 32 (10) (2017) 1067–1074, https://doi.org/10.1080/10426914.2017.1279319.
[25] http://www.grantadesign.com/products/data/materialuniverse.html.
[26] https://github.com/sevamoo/SOMPY.
[27] A. Ultsch, U*-Matrix: A Tool to Visualize Clusters in High Dimensional Data, 2003.
[28] https://scikit-learn.org/stable/modules/clustering.html#k-means.
[30] https://global.kyocera.com/fcworld/first/about.html.
[32] http://www-ferp.ucsd.edu/LIB/PROPS/PANOS/c.html.
[33] L. Glasser, Volume-based thermoelasticity: thermal expansion coefficients and the Grüneisen ratio, J. Phys. Chem. Solids 73 (2012) 139–141, https://doi.org/10.1016/j.jpcs.2011.10.008.
[34] D. Freedman, R. Pisani, R. Purves. Statistics (International Student Edition), 4th, WW Norton & Company, New York, 2007.
[35] W. Jones, N.H. March. Theoretical Solid State Physics, Dover Publications, Inc., New York, 1985.