12-21-2016

Machine Learning Database for Bulk Metallic Glasses

David Robinson
Washington University in St. Louis

Katharine Flores
Washington University in St. Louis

David Robinson

Independent Study

Dr. Katharine Flores

Mechanical Engineering & Materials Science

Machine Learning Database for Bulk Metallic Glasses

Introduction

One of the biggest obstacles materials scientists have encountered in exploring the properties and uses of metallic glass is the ability to predict what alloys have high glass-forming ability (GFA). Since their discovery in the 1960s, there have been many studies and papers published about the GFA of a given alloy or group of alloys. Many studies feature GFA-measuring parameters determined from various chemical, thermodynamic, and physical properties of the constituent elements that compose a given alloy. However, despite repeated attempts to identify universally applicable GFA parameters, there is still no single parameter capable of accurately predicting the GFA of an alloy. This all leads to the question: How can we best identify new alloys to test if we don't have reliable predictors for their GFA? We believe that with a sufficiently large database of alloys, a machine-learning algorithm can be used to identify new candidate alloys. This semester, I was asked to continue the development of such a database, which was started over the summer of this year by another student.

Machine Learning

Machine-learning algorithms predict outcomes by comparing the given input to a large database of input/output pairs, based on an underlying assumption that similar input parameters should yield similar outputs. In order for a machine-learning algorithm to be used for alloy GFA predictions, we need to provide the program with a large database of

alloys with known outcomes. Specifically, we must build a database composed of alloys we already know, including both those with higher GFA (pure metallic glasses) and those with lower GFA (such as metallic glass composites). For this database, known glass-forming alloys will be used as positive results, and known composite-forming alloys will be used as negative results. Because the GFA of an alloy isn't determined solely by its composition, it is necessary to include metadata in our database including but not limited to casting sizes, glass transition temperature, and critical cooling rates. We also intend to include basic mechanical properties such as Young's Modulus, yield strength, and fracture strength. Since this database is still in progress, the specific number and type of parameters we are logging is dynamic and subject to change.

The Database: Where I started

Another student constructed the initial database during the summer of this year. She was tasked with identifying and recording the composition, size, cooling rate, and Glass Transition Temperature for as many alloys as she could find. She created an excel spreadsheet with more than 1,000 alloys with varying amounts of information (some studies did not provide every piece of data she was looking for). Metallic elements are listed across the top of the spreadsheet. For a given alloy (row) the composition was recorded by listing the atomic mass percentage of each element, and blank for elements not contained in the alloy. Cells were also left blank in the event a piece of data was not reported for an alloy. Note, this procedure resulted in two types of blank cells: some are actually zeros, while others are left blank due to missing data. She also recorded bibliographical information for each alloy as well; authors, title, publication, volume, and publication date each get their own column. The vast majority of the data gathered in this initial attempt was taken from papers focusing on GFA in general, or the GFA of specific alloys or alloy groups, so she was not able to provide any data on mechanical properties.

Enhancing Utility of the Database

After meeting with Professor Roman Garnett, the computer science expert who will be helping develop the machine-learning algorithm, it became clear that I needed to adjust the format of the database. Programs such as MATLAB have a hard time both interpreting data sets that are mixed textual and numeric data, as well as dealing with blank cells. Thus my first task was to edit and reformat the database so that it could be exported as a Comma-Separated Value (CSV) file, which is easily manipulated by a variety of programs. Initially, blank cells represented both zeros as well as missing data. To differentiate between the two, I entered '0' for the cells that were true zero-values (like for elements not present in an alloy) and 'nan' for cells with missing data. The cells with mixed textual and numeric data were limited to the size, cooling rate, and Glass Transition Temperature fields. Some cells contained only numbers (e.g. 7) and some contained numbers with units (e.g. 7mm). I had to manually navigate through the spreadsheet to remove the units that sometimes accompanied the numeric data, and instead list them at the top of their respective columns. Finally, I opted to separate the bibliographical information from the actual data in order to make it easier to export the database for use in other programs. I provided each alloy a unique numerical identifier, and copied the bibliographical information into a separate sheet within the same excel file. With each individual source, the alloy identification numbers of the alloy(s) associated with it are listed. This makes it easier to both a) find the source associated with a given alloy and b) associate multiple sources with a single alloy, which will eventually become necessary as I find multiple papers about the same alloy.

Expansion Efforts

Not only was I tasked with editing and reformatting the database, but I am also working on expanding it, primarily by adding data on mechanical properties. I first wanted to fill in the missing data for the alloys already present, so I started by backtracking through the papers cited by the first student. This strategy was unsuccessful, because those sources were primarily about GFA and thermochemistry, with no data on mechanical properties. I

then attempted to find studies conducted on alloys already in the database, but that approach quickly proved unworkable. Searching in Google Scholar or the WUSTL Library system with terms like "$Be_{20}Ti_{16.5}Ni_{9.75}Cu_{15.25}Zr_{38.5}$" fails to yield useful results. In the end, I concluded that the best way for me to add more data to the spreadsheet was to simply add new alloys, and do so by searching for papers that focus on gathering mechanical and physical data. General search terms like "mechanical properties metallic glass" yield tens of thousands of results, the majority of which are relevant to my research. Many of the mechanical property studies focus on a single alloy or alloy group, unlike the publications on GFA, which often compare a much larger number of alloys. This results in fewer alloys to be entered per paper cited, but more consistency in the amount and types of data available.

Current Status and Future Work

The database currently stands at approximately 1,500 alloys, and it will quickly grow as I continue to add new data. The database is easily exportable in a number of formats, which makes it easier to manipulate it with different programs. This opens up potential for us to run the database through a number of different algorithms in a number of different programs without having to reformat the database itself. By separating the bibliographical information from the actual data, there is now greater flexibility and functionality within the database, as it's possible to link multiple sources to a single alloy. I will need to use multiple sources to complete the entries for many of the alloys; so being able to link multiple sources to a single alloy is critical.

Navigating through the database is still very tedious. If you don't already know the alloy's unique identification number, you have to manually find the row with the correct atomic percentage data. For a 5-element alloy, this process can be downright impossible. This makes interacting with the database tedious at best, whether the user is simply searching through the database or is trying to edit it. If I find a new paper on the mechanical properties of a given alloy, I have no efficient way to check the database to see if I already have an entry for it other than by trying to match its composition to that of another alloy I

already have entered. This takes a very long time, and the risk of human error is significant, especially for alloys with more than three or four elements. For the time being, it's fairly safe to assume that I won't be making duplicate entries due to the small size of the database compared to the number of potential alloys. Before long, though, that will no longer be the case. I am working with another student to develop a Python script that will allow us to search the database by any data field (composition, yield strength, glass transition temperature) and will return the row(s) number containing any/all of the alloys that match the query. This will allow users to quickly and efficiently navigate and edit the database, and will make checking for duplicate entries a simple and fast process.

Now that the format of the database is largely stabilized, my remaining jobs are to complete the existing entries and to add as many new ones as I can. I currently have more than 30 papers on my computer awaiting entry, and as soon as I receive the bibliographical information for the fledgling Bulk Metallic Glass Composite database, I will add that as well. Once the database is sufficiently populated, I plan to work with Professor Garnett on developing and testing the machine-learning algorithm. Ideally, once the algorithm(s) are up and running, I will be able to test the alloy combinations it comes up with, and use those results to further refine the algorithms.