Optimizing Downsampling in Variable Density Experimental Data: Predicting Metallic Glasses

Brenna M Gibbons[1]
brennamg@stanford.edu

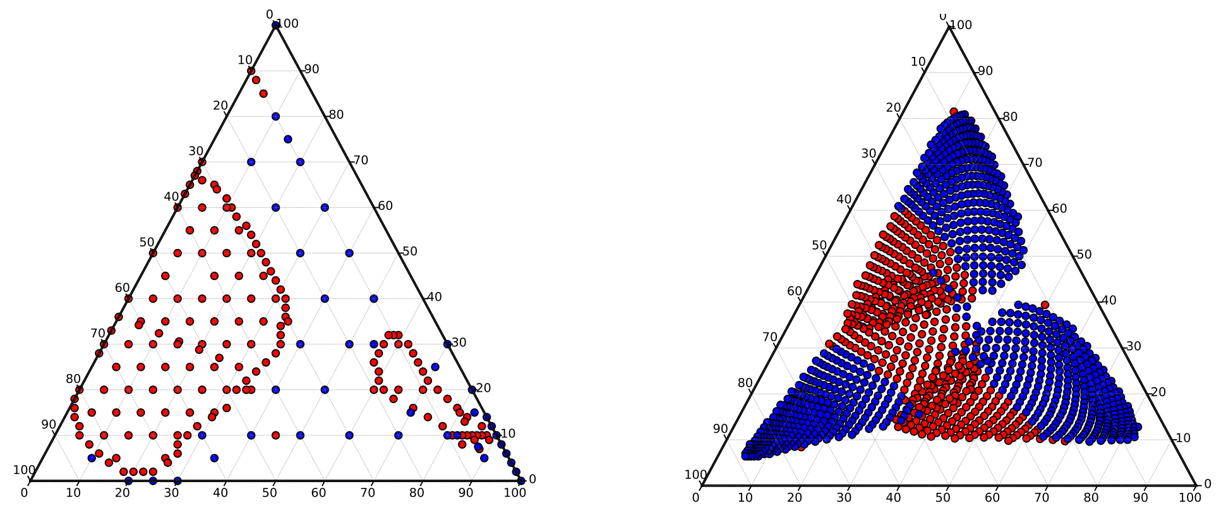
Cooper W Elsworth[2,3]
coopere@stanford.edu

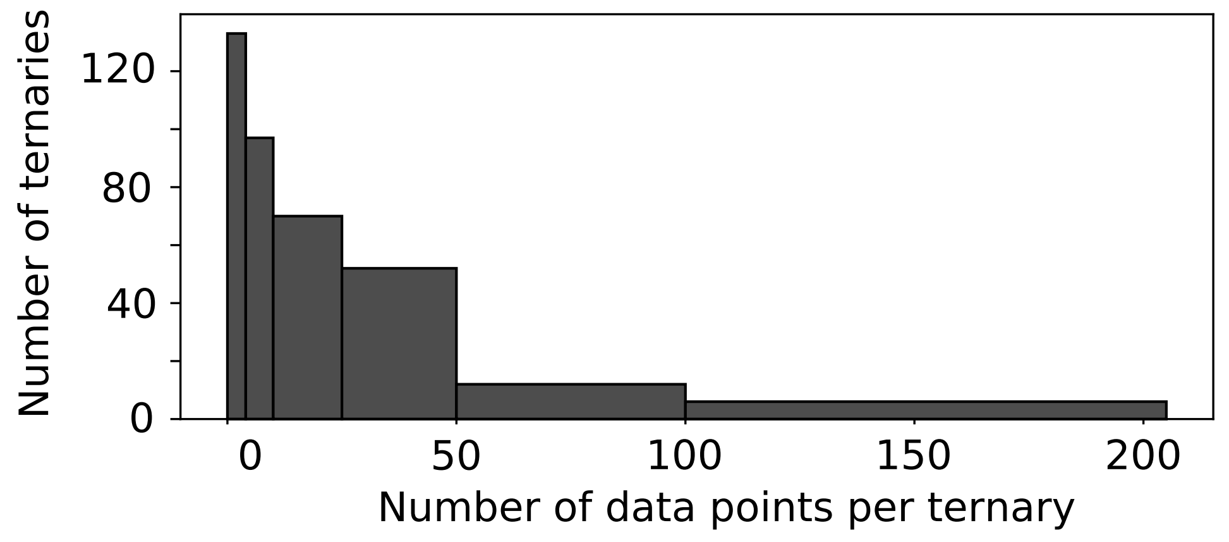
[1]Department of Material Science and Engineering
[2]Department of Geophysics
[3]Institute for Computational & Mathematical Engineering

Metallic glasses are a unique class of materials that combine many of the desirable properties of crystalline metals, such as good electrical conductivity, with the advantages of amorphous glasses, such as ease of processing and high resistance to corrosion. However, there is no analytical formula for determining whether an arbitrary alloy composition is capable of forming a metallic glass, and such predictions are difficult using empirical models[1]. We demonstrate the capability of machine learning algorithms to predict the glass-forming ability of ternary alloys for which the models have no prior information. Additionally, we explore how downsampling affects the model performance when adding dense data to a sparse dataset.

Data

The data is based on the Landold-Bornstein dataset, which contains ~5700 data points over ~300 ternary alloys [Ward et al., 2016]. We also combine this relatively sparse dataset with dense data from high-throughput experiments on 9 ternaries with ~1300 points each[3].





Models

Logistic Regression

An unweighted logistic regression model implemented using sklearn[5].

Support Vector Machine

A support vector machine model implemented using sklearn[5] with an RBF kernel.

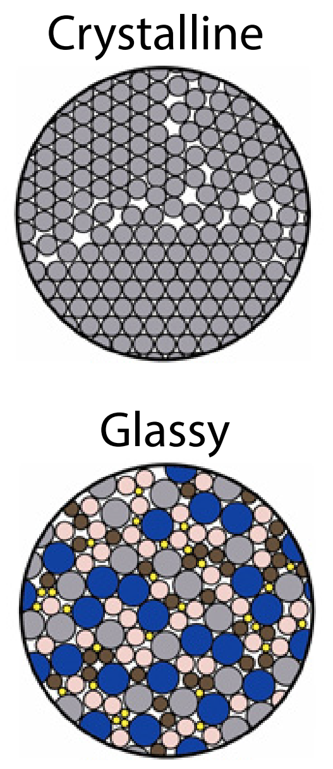
Random Forest

An ensemble of decision trees implemented using sklearn[5], with no maximum depth, 500 trees, and sqrt(n) features considered at each split.

Neural Network

A simple neural network with one hidden layer of 100 neurons and a relu activation function.

Note: all features were used for each model. The same 20% of the sparse dataset (1077 points) was used as a dev set for each model. For the table shown to the right, the training set was 4290 points from the sparse dataset.



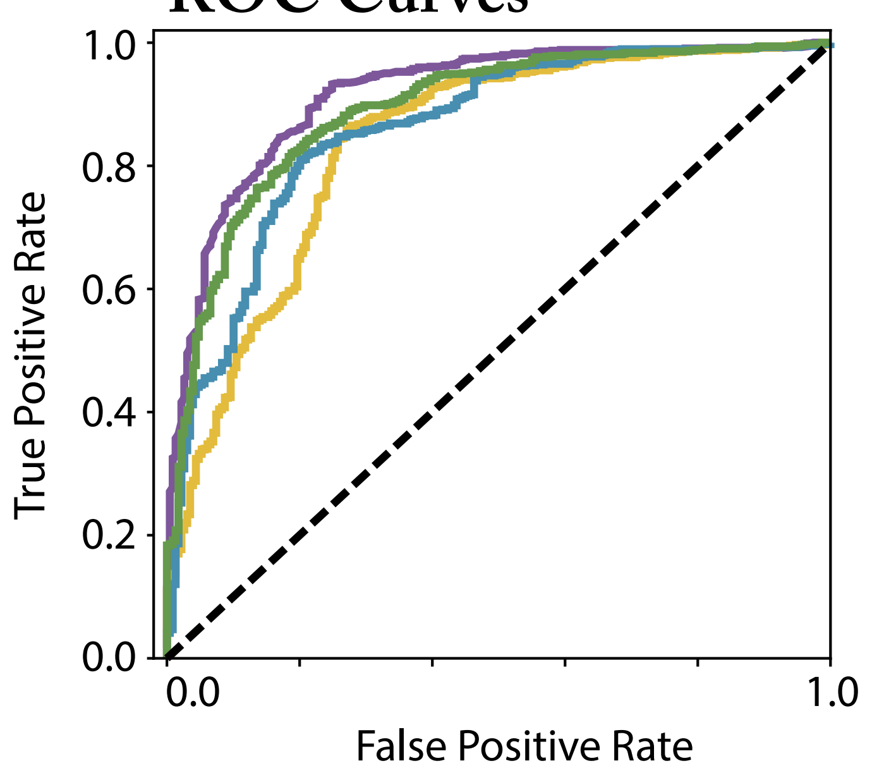
Features

The raw data from Ward et al., 2016[2] contains a string representation of a composition (e.g. "Al80Ni15Zr5") and a classification of "0" for glass and "1" for crystalline. The data was featurized using the Materials-Agnostic Platform for Informatics and Exploration (Magpie)[4]. 51 features represent the composition of each possible element considered, while 144 additional features capture the weighted average, minimum, maximum, and standard deviation of various physical parameters (e.g. the mean electronegativity, weighted by the composition).

Results

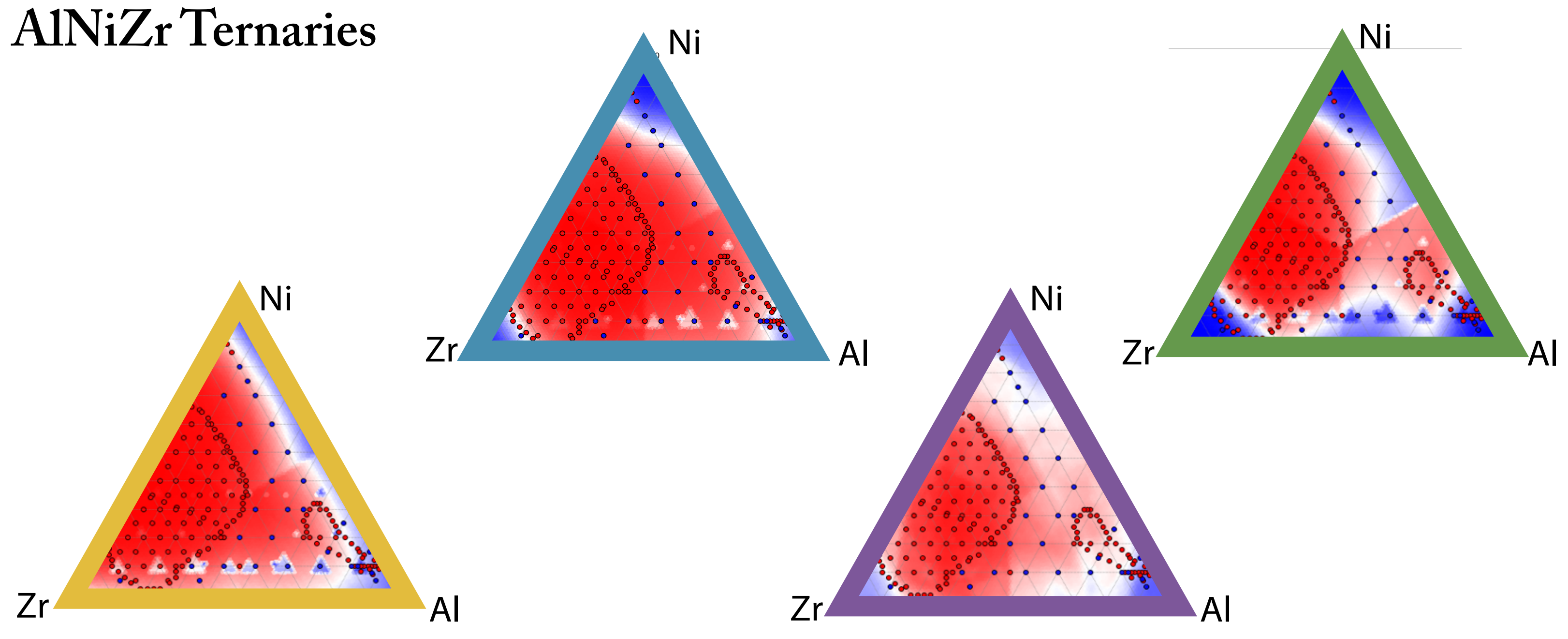
For the metrics we chose, the random forest model and the neural network model performed the best on the sparse dataset. Adding the high throughput data improved the performance of the random forest model, even at low sampling rates.

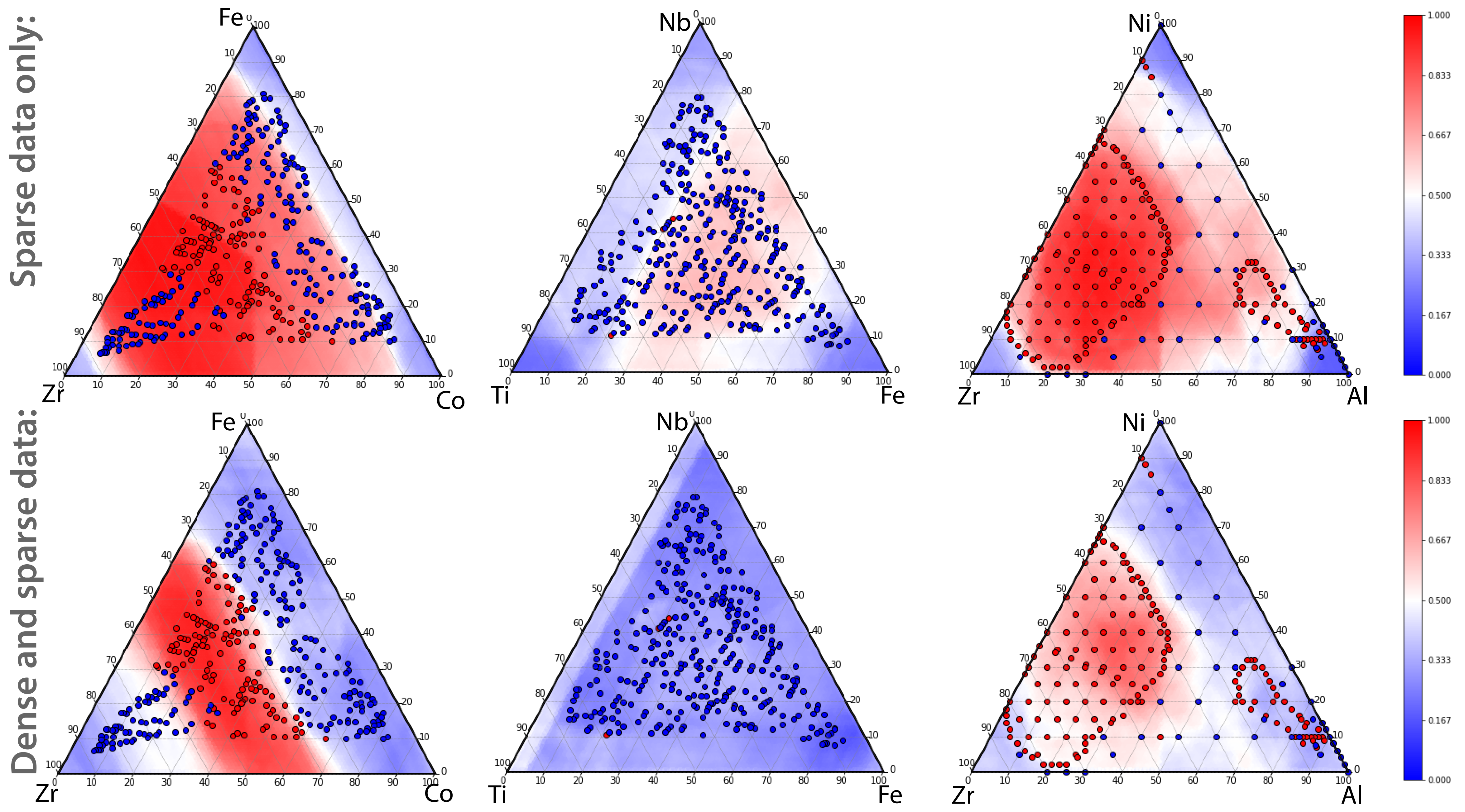
ROC Curves



	Logistic Regression	SVM	Random Forest	Neural Network
Train Accuracy	82.4%	84.1%	**100%**	86.6%
Dev Accuracy	85.7%	85.8%	**88.3%**	84.7%
AUROC	0.836	0.859	**0.915**	0.876
Log-Loss	0.384	0.372	**0.313**	0.336

AlNiZr Ternaries

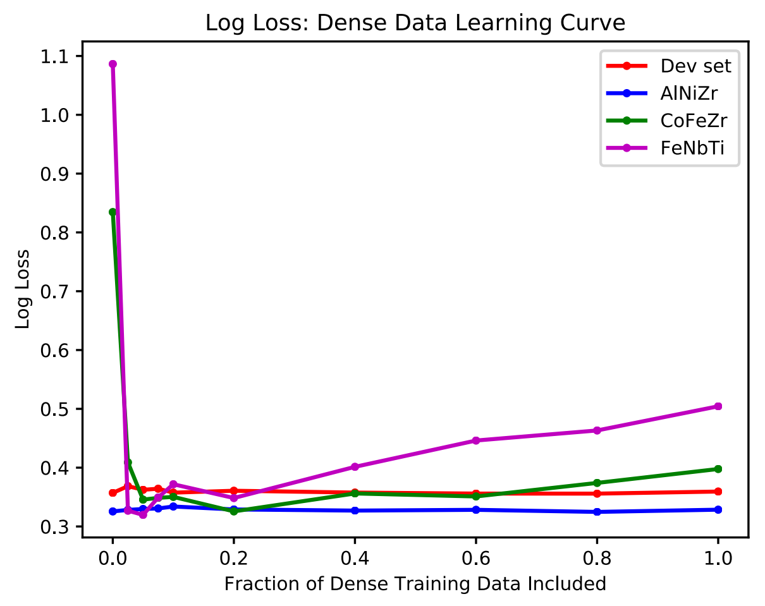


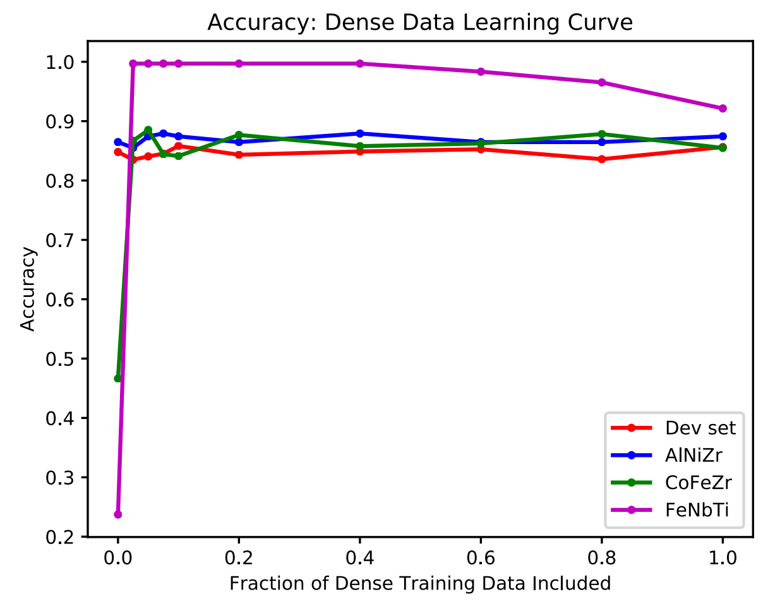


Discussion

The models we constructing had similar performances to the paper in which we found the data [2]. The most obvious way to improve this model would be to add more data, particularly data from ternaries that aren't represented here; however, it may not be possible to reach perfect accuracy, since the experimental results occasionally disagree.

One surprising result of our investigation was how little of the dense data is required to achieve a noticeable improvement in performance when predicting on specific ternaries. This result is notable because these experiments are performed at SLAC, where time is extremely limited, so any reduction in data collection time is valuable. Moving forward, we hope that these models can be used to guide the discovery of new metallic glasses.





Future Work

The next step for this project would be to further tune the parameters of each model. In addition, we would like to explore different methods of downsampling, for example, taking more points near phase boundaries.

References

[1] Li, Y.; Zhao, S.; Liu, Y.; Gong, P.; and Schroers, J. (2017) *ACS Combinatorial Science*.
[2] Ward, L.; Agrawal, A.; Choudhary, A.; and Wolverton, C. (2016) *npj Computational Materials 2*.
[3] Ren, F.; Pandolfi, R.; Van Campen, D.; Hexemer, A.; Mehta, A. (2017) *ACS Combinatorial Science*.
[4] https://bitbucket.org/wolverton/magpie
[5] http://scikitlearn.org/