# Predicting Protein Relationships Using Machine Learning

Eli Blaney and Soochin Cho

*Departments of Biology, Creighton University, Omaha, NE*

## Abstract

Protein-protein interactions have been discovered and digitally collected over time, giving rise to large collections of protein data [1]. Machine learning algorithms can be fed structural and relational information about pairs of proteins to predict the strength of arbitrary protein relationships from their structural data. However, various machine learning approaches can perform very differently, and inefficiencies can cause significant setbacks in research. By assessing the viability of several classification algorithms trained on this protein data, each algorithm can be studied to find potential enhancements that can lead to more accurate and consistent predictions.

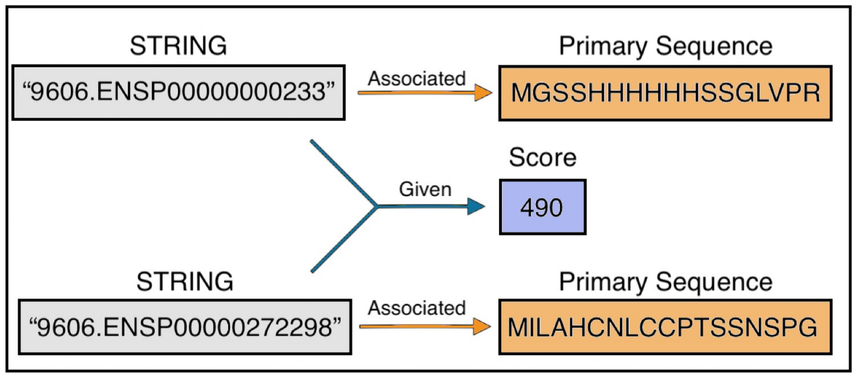


**Figure 1:** Association created between the STRING ID for a protein and its primary sequence. Pairs of STRING IDs are given scores in the STRING database which are used to train the models.

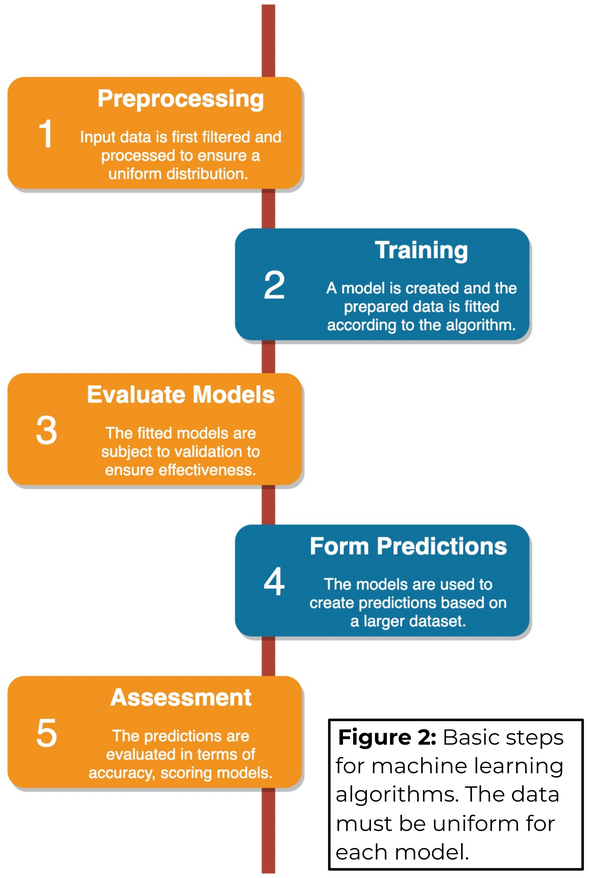


**Figure 2:** Basic steps for machine learning algorithms. The data must be uniform for each model.

## Background

- Machine learning algorithms can be fed structural and relational information about pairs of proteins (Figure 1), allowing them to predict the relationship between new pairs of proteins [2].
- Many machine learning algorithms can be applied to analyze this type of data, each with its own pros and cons. In my study, the viability of four algorithms were assessed on the ability to predict protein-protein functional relationships.
- Comparing the viability of multiple approaches is crucial toward developing stable computational protein analysis techniques.
- Measuring the effectiveness of predicting functional association scores allows one to better decide between various machine learning algorithms in particular protein contexts.

## Model Design

The code was written using the Python (v3.9) programming language, employing the scikit-learn library to create the machine learning models [3]. Two classification and two regression models were chosen:

- A **support vector classifier**, which uses high-dimensional feature spaces
- A **random forests classifier**, which uses an ensemble of decision trees
- A **Bayesian ridge regression**, which uses Bayes' theorem
- A **logistic regression**, which uses a one-vs-rest scheme

## Results & Discussion

- All but the Bayesian ridge regression could predict relationships more than half correctly.
- Using the data in Figure 3, it is apparent that the support vector classifier slightly outperformed when predicting the classes of the functional relationship.
- In the context of using these models, the four algorithms appear to require some further adjustments and more detailed data before being used to accurately and consistently predict the relationships between proteins.
- Scripts and data can be found on this project's GitHub repository: `github.com/eliblaney/pdb-learn`

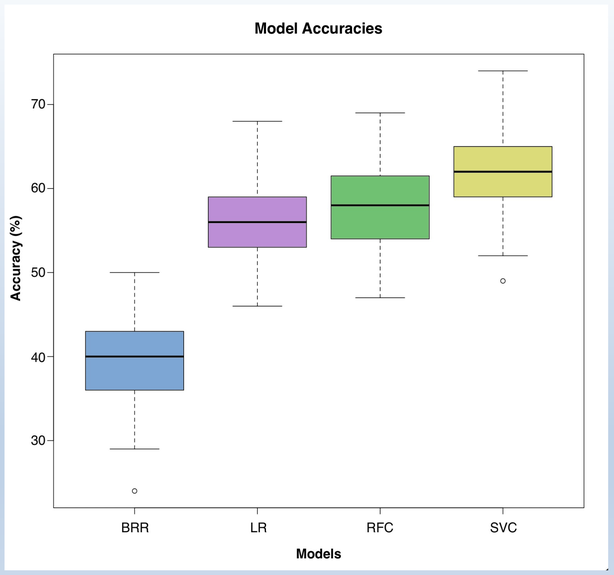


**Figure 3:** Box plot for the model class-based accuracies for Bayesian ridge regression (BRR), logistic regression (LR), random forest classifier (RFC), and support vector classifier (SVC). n = 10,000.

## Future Research

Using another database, I am expanding the inputs of the algorithms to include ligand data, binding affinity, and the 3D atomic arrangement of each protein to create more powerful models. High memory utilization is an issue that I have been working on overcoming as I approach this goal.

### Acknowledgements

Barry Goldwater Scholarship & Excellence Education Foundation
Creighton University Center for Undergraduate Research and Scholarship
Creighton University Honors Program

### References

[1] Coluzza, I., Journal of Physics: Condensed Matter. **2017**, (29), no. 14.
[2] Szklarczyk, D., et al., Nucleic Acids Research. **2019**, (47), no. D1.
[3] Pedregosa, F., et al., The Journal of Machine Learning Research. **2011**, (12).