A Machine Learning Approach to Optimizing the Tensile Strength of Starch-Based Bioplastics

Grant Proposal

Sriaditya Vaddadi

Massachusetts Academy of Math & Science at Worcester Polytechnic Institute

Worcester, MA

Table of Contents

Abstract (RQ) or Executive Summary (Eng)

Plastics are used all over the world, and there has been a boom in the need for plastics due to the growing infrastructure and technologies of the modern era. However, current synthetic plastics come at a great environmental cost. One of the biggest problems with synthetic plastics, or plastics made from petroleum, is that plastic production accounts for a significant amount of carbon dioxide emissions, which may lead to global warming or climate change. As a result of this, many scientists have begun research on bioplastics, a type of plastic that is created using organic materials, such as starch. These bioplastics are generally biodegradable, meaning that landfills will no longer need to contain large amounts of plastics that take thousands of years to degrade. Furthermore, the production of bioplastics emits much less carbon dioxide into the atmosphere as they do not need oil/petroleum to be fabricated. However, there are also many problems currently haunting the field of bioplastics. One of the most glaring problems with bioplastics is that they are, on average, much weaker than their synthetic counterparts. As a result, this project wishes to use Machine Learning (and specifically, regression algorithms) to estimate and optimize the tensile strength of starch-based bioplastics. The independent variables used were a subset of the amylose, amylopectin (two components of starch), and glycerol (plasticizer) concentrations. The model was run for 250 iterations, each iteration randomly generating a 70/30 training/verifying split. Overall, a polynomial regression model with the independent variables of amylose and glycerol concentrations performed the strongest, with an R-squared of roughly 0.484 and an RMSE of roughly 2.606.

Keywords: Bioplastics, Starch-Based, Machine Learning, Regression, Amylose, Glycerol, Polynomial Regression

A Machine Learning Approach to Optimizing the Tensile Strength of Starch-Based Bioplastics

Currently, synthetic plastics are one of the most harmful yet constantly used materials in the world. From being one of the largest factors in carbon dioxide emissions (accounting for roughly 3.3% of all CO2 emissions), to the rise in microplastics causing many severe health issues in humans, plastic production has harmed and will continue to harm the lives of many organisms, including humans (Ritchie, 2023; Pilapitiya & Ratnayake, 2024). These toxic microplastics can be found in all sorts of everyday actions, such as brushing your teeth or even breathing (Li et al., 2023). Similarly, another large problem caused by plastic production is their carbon dioxide emissions. Therefore, there have been many climate concerns regarding plastic production. As the climate change situation is becoming more dire every day, there are rising tensions about the future of life on Earth. If we continue the way we are now, natural disasters will increase in severity, the ice caps may continue to melt even further, and the overall temperatures will continue to rise (NASA). Specifically, the global temperature will likely rise by roughly 1.5 degrees Celsius (or an extra 2.7 degrees Fahrenheit) over the next two decades (IPCC). While such a change seems minimal, it is important to note that such a warming will be on average, and that the extremities in weather will continue to get worse and worse. Due to all these staggeringly dangerous factors, scientists are working to combat the rising levels of carbon dioxide in the atmosphere.

One way in which scientists are aiming to combat global carbon emissions is the field of bioplastics, which uses natural, biodegradable materials to create bioplastics that can replace current synthetic plastics. According to a review article by Nanda et al. (2022), there were roughly 2.2 million tons of bioplastics produced in 2021, with 850,000 metric tons of non-biodegradable bioplastics and 1.35 million metric tons of biodegradable bioplastics. Although this is great progress, these numbers are nowhere close to the almost 370 million tons of synthetic plastics produced in the same year, as shown in Figure 1 (Nanda et al., 2022). Some specific benefits of bioplastics are that they are usually biodegradable and that there are very low carbon dioxide emissions associated with the production of

bioplastics (Nanda et al., 2022). This is particularly useful since a switch to just biodegradable bioplastics would not only lessen the carbon dioxide emissions by a decent amount, but it would also cause landfills to clear up and stop expanding. For all the benefits of bioplastics, however, there are a few problems. For example, starch-based bioplastics are generally weaker than their synthetic counterparts (Abe et al., 2021).

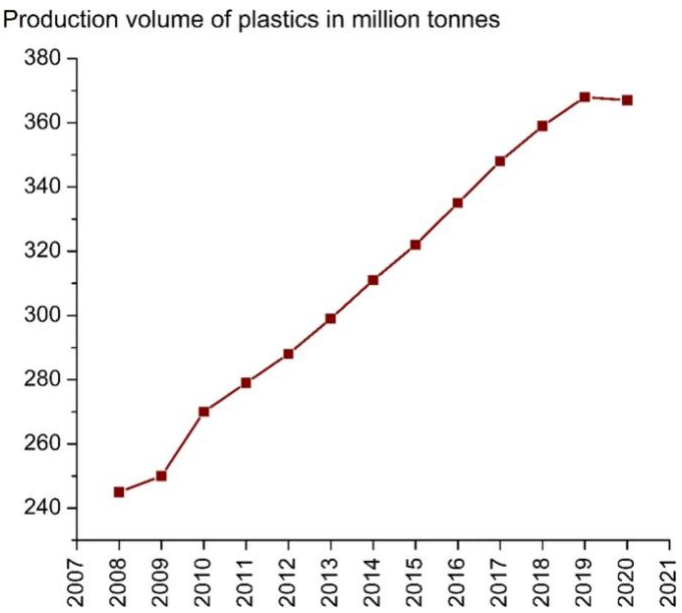

Figure 1: The production volume of plastics in millions of tons. We can see how there has been a constant rise in how much plastic has been produced throughout the years (Nanda et al., 2022).

Since this problem is so prevalent in the field of bioplastics, there have already been many attempts to fabricate stronger bioplastics. Some examples of studies that have tried to fabricate starch-based bioplastics have looked at the starch from foods such as corn, rice, and jackfruit (Marichelvam et al., 2019; Nguyen et al., 2022). However, it is unclear in these articles if the materials and methods used by the authors were the optimal materials and methods to create the strongest possible bioplastic. For example, Figure 2 shows that the researchers stopped testing even though the tensile strength of the bioplastic never began to decrease, meaning that an optimal tensile strength may not have been found (Nguyen et al., 2022).

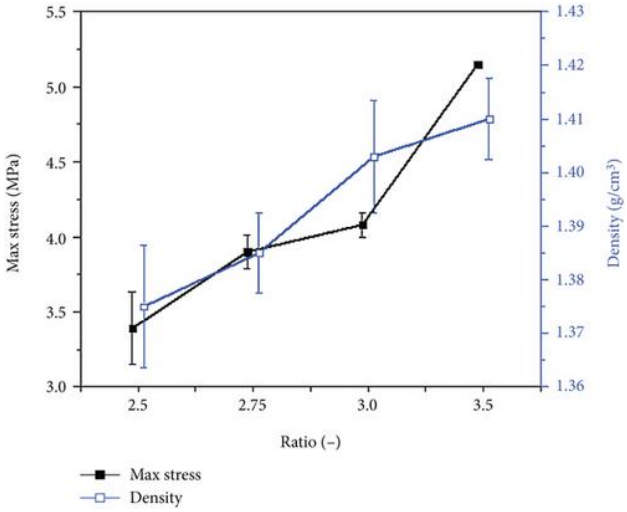

Figure 2: How the Starch: Glycerol Ratio of Starch-Based Bioplastics affects the Max stress of Bioplastics (Nguyen et al., 2022).

Therefore, more research is necessary to find how strong these starch-based bioplastics can get. An interesting idea that is being talked about in the field of bioplastics currently is using machine learning to assist in the fabrication of bioplastics. For example, in an article by Kuenneth et al. (2022), neural networks were used to generate bioplastics that were the closest to pre-existing synthetic plastics. A review paper by Guardia et al. (2024) showed that, currently, there are many different types of machine learning algorithms being used to fabricate bioplastics. Specifically, the article outlines how Supervised Learning algorithms (both Regression and Classification) are being used in the field currently. These Supervised Learning algorithms are also being used in tandem with Unsupervised Learning algorithms (Guardia et al., 2024). This proposal aims to similarly work with ML to make the bioplastic fabrication process much more efficient. The overall goal of the proposal is to make a machine learning model (using regression algorithms) that can take in the amylose, amylopectin, and glycerol concentrations of a bioplastic and then output the tensile strength of the bioplastic. Pre-existing data from other scientific papers that focused on fabricating starch-based bioplastics with glycerol will be used as training data for this model. Once the model is sufficiently trained, other pre-existing data will be used to test/train the model and make any necessary changes. After a few iterations of this kind, the final model should be able to accurately predict the tensile strength of any starch-based bioplastic given the bioplastic's amylose, amylopectin, and glycerol concentrations. A regression equation can then be further used to find the optimal values of amylose, amylopectin, and glycerol concentrations that would maximize the tensile strength of the bioplastic. The model made in this proposal should help those who wish to fabricate bioplastics by letting them find the optimal starch and glycerol concentrations to optimize tensile strength. Further research in the field could consider other factors used to fabricate bioplastics, such as different plasticizers like sorbitol, as well as other properties of bioplastics (such as elongation or water resistance) to optimize.

Section II: Specific Aims

This proposal's objective is to make a machine learning model that will take the inputs of the amylose, amylopectin, and glycerol concentrations of a bioplastic and should output the expected tensile strength of the plastic. The central hypothesis of this project is that as the concentration of starch decreases, or as the concentration of glycerol increases, the tensile strength of the bioplastic will decrease and that a regression algorithm can model these correlations.

Specific Aim 1: The first, main goal of this proposal is to create a machine learning model that can be used to find a regression equation that correlates a subset of the amylose, amylopectin, and glycerol concentrations of a bioplastic to said bioplastic's tensile strength.

Specific Aim 2: Once this model is fully operating, the second goal of this proposal is to test out the model against other pre-existing data that was not used to train the model and then use the results to tweak the model as necessary.

Specific Aim 3: A third goal for this proposal is to see what type of regression algorithm would, in general, best suit the process of estimating the mechanical properties of a bioplastic given the independent variables previously mentioned.

The expected outcome of this work is to have a fully operational model that can be used in the field of bioplastics to save time and energy when testing for tensile strengths, thereby making work in the field more efficient, and incentivizing others to focus on different mechanical properties of bioplastics, such as elongation at break or water resistance.

Section III: Project Goals and Methodology

This proposal is quite significant, as once machine learning can be used to maximize the tensile strength of starch-based bioplastics, researchers can then begin to optimize other mechanical properties of bioplastics. Maximizing the tensile strength of bioplastics is very important since one of the main problems with bioplastics is their strength. However, tensile strength is not the only important factor of

bioplastics: mechanical properties such as elongation and water resistance can also be improved. If a bioplastic with maximized mechanical properties is fabricated, there is a very high likelihood that current synthetic plastics can be replaced with these bioplastics, which would not only be biodegradable but would also reduce the harm of microplastics. Currently, there are little to no machine learning models being used to maximize tensile strength. Some papers, such as those by Hendrawan et al. (2020), used a method known as the Response Surface Method to maximize the ultimate tensile strength of Arrowroot starch-based bioplastic. However, this paper only focuses on a specific type of starch, and the Response Surface Method is quite different from the methodology used in this proposal. Therefore, it is fair to say that this proposal is very innovative in how it tackles the topic of maximizing the tensile strength of starch-based bioplastics. This paper proposes to use a machine learning model (and specifically, a regression algorithm) to find the optimal amylose, amylopectin, and glycerol concentrations to maximize the tensile strength of starch-based bioplastics. The data used to test the model will come from many other scientific articles that fabricated starch-based bioplastics with glycerol as a plasticizer since there is not a great dataset to use for this proposal. The collected data was then normalized. The code for the machine learning model was written in Python, and the data was randomly split into a 70/30 training/verifying split over 250 iterations, with the R-squared and RMSE error terms being averaged out over these 250 iterations to see which algorithm had the highest accuracy.

Specific Aim #1:

The first, main goal of this proposal is to create a machine learning model that can be used to find a regression equation that correlates a subset of the amylose, amylopectin, and glycerol concentrations of a bioplastic to said bioplastic's tensile strength. The approach for this aim is to use Python and testing data from other papers that have fabricated starch-based bioplastics with glycerol as a plasticizer in the past. Seventy percent of this data will be used to train the model.

Justification and Feasibility. Our justification for this approach is that since there is not a general dataset for tensile strengths of starch-based bioplastics using glycerol as a plasticizer, it is instead easier to gather training data from multiple different sources. The data was then normalized to account for different lab environments. A machine learning model (especially a regression model) is especially useful in the case of studying data like this, since the goal of this project is to find an equation that correlates the three independent variables to the dependent variable of tensile strength.

Summary of Preliminary Data. The machine learning model is currently being coded in W3Schools' TryIt Editor, but the code should soon be moved to a GitHub repository. Figure 3 shows a snippet of the code. Since there is a machine learning model that can return the tensile strength of a bioplastic given a subset of the bioplastics' amylose, amylopectin, and glycerol concentrations, this aim was successfully reached.

```
for randomStateValue in range (numOfIterations):
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size =
0.3, random_state = random.randrange(0, numOfIterations, 1))

  linearReg = linear_model.LinearRegression()
  linearReg.fit(X_train, y_train)
  y_pred_Linear = linearReg.predict(X_test)
  rsquaredLinear += linearReg.score(X_test, y_test) / numOfIterations
  mseLinear += mean_squared_error(y_test, y_pred_Linear) /
numOfIterations

  polynomialReg = PolynomialFeatures(degree=2)
  X_polynomial = polynomialReg.fit_transform(X_train)
  polynomialReg.fit(X_polynomial, y_train)
  otherLinearReg = linear_model.LinearRegression()
  otherLinearReg.fit(X_polynomial, y_train)
  y_pred_Polynomial =
otherLinearReg.predict(polynomialReg.fit_transform(X_test))
  rsquaredPolynomial +=
otherLinearReg.score(polynomialReg.fit_transform(X_test), y_test) /
numOfIterations
  msePolynomial += mean_squared_error(y_test, y_pred_Polynomial) /
numOfIterations

  supportVectorReg = SVR(kernel = 'rbf')
  supportVectorReg.fit(X_train, y_train)
  y_pred_SVR = supportVectorReg.predict(X_test)
  rsquaredSVR += supportVectorReg.score(X_test, y_test) /
numOfIterations
  mseSVR += mean_squared_error(y_test, y_pred_SVR) / numOfIterations
```

Figure 3: A snippet of the code used to split the data (70/30 split on training/testing) over 250 iterations as well as to train the different regression algorithms.

Expected Outcomes. The overall outcome of this aim is to make a machine learning model that can take in the amylose, amylopectin, and glycerol concentrations of a starch-based bioplastic and return the tensile strength of said bioplastic. This knowledge will be used to help make the bioplastic fabrication process much more efficient (allowing scientists to make a bioplastic as strong as possible without requiring excessive testing). It will also allow for further advancements in the field if the machine learning strategy is to be applied to other mechanical properties of bioplastics.

Potential Pitfalls and Alternative Strategies. We might run into the problem that a specific regression algorithm does not provide useful results (i.e. if linear regression gives a low R-squared value,

there is a good chance that we cannot use linear regression), in which case we will test other regression algorithms.

Specific Aim #2:

Once this model is fully operating, the second goal of this proposal is to test out the model against other pre-existing data that was not used to train the model and then use the results to tweak the model as necessary. Our approach for this aim is to take 30% of the full dataset acquired and verify the model on this data, with there being 250 random iterations of these 70/30 splits.

Justification and Feasibility. Our rationale for this approach is that this is generally how machine learning models are tested. If we find the experimental tensile strength of the test data is not relatively close to the model's output, we will collect more data as necessary until the model is generally consistent.

Summary of Preliminary Data. 250 iterations of random 70/30 training and testing splits were done, and the average R-squared and RMSE values were printed. The most accurate regression algorithm was found to be polynomial regression, with a R-squared value of around 0.484 and a RMSE of around 2.606. More data regarding all the regression algorithms can be found in Specific Aim #3's summary of preliminary data.

Expected Outcomes. The overall outcome of this aim is to reduce the model output's error and to make the model more usable for new research. Since we do not want the model to overfit (match the data too closely) to our training data, we will have multiple iterations of testing. Similarly to the first aim, the reason for this aim is that this model can be used in the future to help accurately predict the tensile strength of a bioplastic given the amylose, amylopectin, and glycerol concentration of the bioplastic.

Potential Pitfalls and Alternative Strategies. If the model was only trained and tested once, the results (i.e. R-squared and RMSE values) would not accurately represent the average accuracy of each regression algorithm. Therefore, the code runs 250 iterations of random 70/30 training/testing splits, with the average R-squared and RMSE values being recorded.

Specific Aim #3:

A third goal for this proposal is to see what type of regression algorithm would, in general, best suit the process of estimating the mechanical properties of a bioplastic given the independent variables previously mentioned. Our approach for this goal is to train and test out multiple different regression algorithms and see which algorithm has the highest accuracy.

Justification and Feasibility. Our rationale for this approach is that if we were to only test one algorithm, it would be left uncertain if our ML model was the most accurate. Therefore, we are testing multiple algorithms.

Summary of Preliminary Data. Figures 4 and 5 show the R-squared and RMSE values respectively of the four regression algorithms being tested currently. As shown in these figures, polynomial regression performs the best, with an R-squared value of roughly 0.484 and a RMSE of roughly 2.606. It is worth noting, however, that this is with the independent variables of just amylose and glycerol concentrations (rather than amylose, amylopectin, and glycerol), since the accuracy of the models improves when the amylose concentration is omitted.

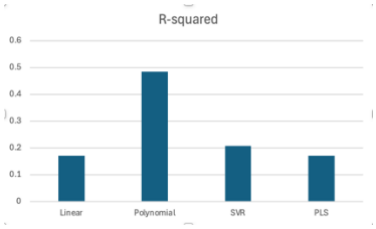

Figure 4: The R-squared values of Linear Regression (~0.17), Polynomial Regression (~0.484), SVR (~0.207) and PLS regression (~0.17)

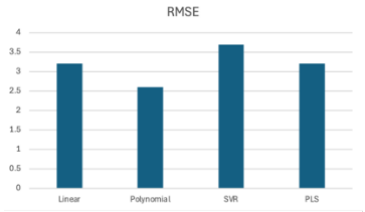

Figure 5: The RMSE of Linear Regression (~3.21), Polynomial Regression (~2.61), SVR (~3.7), and PLS (~3.21)

Expected Outcomes. The overall outcome of this aim is to see which of the regression algorithms we are testing has the highest accuracy. Once we find which regression algorithm has the highest accuracy, we can then use that algorithm to help researchers accurately predict the tensile strength of their bioplastic before even fabricating the plastic.

Potential Pitfalls and Alternative Strategies. Although we are trying many different regression algorithms, we still cannot be sure that the best regression algorithm to model the data is one of the algorithms we are using (i.e. if we only use linear and polynomial regressions, but some other untested algorithm would have a higher accuracy). Therefore, future steps include testing out multiple other regression algorithms and comparing their accuracies with the accuracies of the algorithms we have already tested.

Section IV: Resources/Equipment

For this proposal, a computer with Python will be used to run the machine learning model.

Section V: Ethical Considerations

There will be no tests on living organisms, since the project is just a machine learning model tested on pre-existing data. Therefore, there are no ethical considerations necessary.

Section VI: Timeline

Phase 1a. Continue to read Journal Articles/Patents in the field of ML with regards to bioplastic: Oct 29th to Nov 18th

Phase 1b. Look at previous Data Sets used in the field: Oct 29th to Nov 18th

Phase 2a. Begin using Pre-Existing data to form a data set on which the regression model can be trained: Nov 11th to Nov 25th

Phase 2a, i. Decide on a type of regression model to use: Nov 11th to Nov 14th

Phase 2a, ii. Use data points from bioplastics papers that would fit this model: Nov 14th to Nov 25th

Phase 3a. Create Regression Model and test it on Data Set created in Phase 2a: Nov 25th to Dec 2nd

Phase 3a, i. Get Data in a Usable Form to make the Regression (i.e. Excel or Python): Nov 25th to Nov 28th

Phase 3a, ii. Run the regression algorithm on the dataset: Nov 25th to Dec 2nd

Phase 3b. Prepare for December Fair: Nov 25th to Dec 9th

Phase 3b, i. Make Poster for December Fair: Dec 3rd to Dec 4th

Phase 3b, ii. Make and Practice Pitch for December Fair: Dec 5th to Dec 6th

Phase 4a. Update the Data Set found in Phase 2 to add more data: Dec 9th to Dec 20th

Phase 4b. Rerun the regression algorithm using the data set from 4a: Dec 20th to Dec 24th

Section VII: Appendix

Section VIII: References

Abe, M. M., Martins, J. R., Sanvezzo, P. B., Macedo, J. V., Branciforti, M. C., Halley, P., Botaro V. R., Brienzo, M. (2021). Advantages and Disadvantages of Bioplastics Production from Starch and Lignocellulosic Components. *Polymers (Basel), 13*(15). 10.3390/polym13152484

Guardia, C., Caseiro, J., Pires, A. (2024). Machine learning to enhance sustainable plastics: A review. *Journal of Cleaner Production, 474*. https://doi.org/10.1016/j.jclepro.2024.143602.

Hendrawan, Y., Putranto, A. W., Fauziah, T. R., Argo, B. D. (2020). Modeling and Optimization of Tensile Strength of Arrowroot Bioplastic Using Response Surface Method. *IOP Conference Series: Earth and Environmental Science, 515.* https://doi.org/10.1088/1755-1315/515/1/012079

IPCC. (2021, August 9). *Climate change widespread, rapid, and intensifying – IPCC.* https://www.ipcc.ch/2021/08/09/ar6-wg1-20210809-pr/

Kuenneth, C., Lalonde, J., Marrone, B. L., Iverson, C. N., Ramprasad, R., & Pilania, G. (2022). Bioplastic design using multitask deep neural networks. *Communications Networks, 3*(96). https://doi.org/10.1038/s43246-022-00319-2

Li, Y., Tao, L., Wang, Q., Wang, F., Li, G., & Song, M. (2023). Potential Health Impact of Microplastics: A Review of Environmental Distribution, Human Exposure, and Toxic Effects. *Environmental Health, 1*(4). https://doi.org/10.1021/envhealth.3c00052.

Marichelvam, M. K., Jawaid, M., & Asim, M. (2019). Corn and Rice Starch-Based Bio-Plastics as Alternative Packaging Materials. *Fibers, 7*(4). **https://doi.org/10.3390/fib7040032.**

NASA. (n. d.). *The Effects of Climate Change.* https://science.nasa.gov/climate-change/effects/

Nguyen, T. K., That, N. T. T., Nguyen, N. T., & Nguyen, H. T. (2022). Development of Starch-Based Bioplastic from Jackfruit Seed. *Advances in Polymer Technology, 2022*(1). https://doi.org/10.1155/2022/6547461.

Pilapitiya, P. G. C. N. T., & Ratnayake, A. S. (2024). The world of plastic waste: A review. *Cleaner Materials, 11*. https://doi.org/10.1016/j.clema.2024.100220.

Ritchie, H. (2023, October 5). *How much of global greenhouse gas emissions come from plastics?* Our World in Data. https://ourworldindata.org/ghg-emissions-plastics

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*. https://doi.org/10.1007/s42979-021-00592-x