JOURNAL OF
EMERGING INVESTIGATORS

Article

Machine learning-based enzyme engineering of PETase for improved efficiency in plastic degradation

Arjun Gupta[1], Sangeeta Agrawal[2]

[1] Augustus Academy, Scottsdale, Arizona

[2] Wright State University, Dayton, Ohio

SUMMARY

Globally, nearly one million plastic bottles are produced every minute. These non-biodegradable plastic products are composed of polyethylene terephthalate (PET). In 2016, researchers discovered PETase, an enzyme from the bacteria *Ideonella sakaiensis* that breaks down PET and nonbiodegradable plastic. Temperatures above 60 – 65 °C are optimal for PET degradation as the polymer chain fluctuates in this range, allowing water molecules to enter and weaken the chains. However, PETase has low efficiency at these temperatures, thus limiting its usage. Here, we optimized the rate of PET degradation by PETase by designing new mutant enzymes that could break down PET much faster than PETase. We used machine learning-guided directed evolution to modify PETase to have a higher optimal temperature (T_{opt}), which would allow the enzyme to degrade PET more efficiently. First, we trained three machine learning models to predict T_{opt} with high performance, including Logistic Regression, Linear Regression, and Random Forest. We then used Random Forest to perform machine learning-guided directed evolution. Our algorithm generated hundreds of mutants of PETase and screened them using Random Forest to select mutants with the highest Topt. After 1000 iterations, we produced a new mutant of PETase with T_{opt} of 71.38 °C. We also produced a new mutant enzyme after 29 iterations with Topt of 61.3 °C. To ensure these mutant enzymes would remain stable, we predicted their melting temperatures using an external predictor and found the 29-iteration mutant had improved thermostability over PETase. Using this approach and novel algorithm, scientists can optimize additional enzymes for improved efficiency.

INTRODUCTION

According to the United Nations, over 200 million tons of plastic are produced every year, and 91% of all plastic produced is not recycled (1, 2). One non-polluting recycling and waste management method for plastic is enzymatic recycling. In 2016, researchers in Japan identified a bacterium, *Ideonella sakaiensis*, that consumed and successfully degraded polyethylene terephthalate (PET), the most common form of non-biodegradable plastic. Yoshida *et al.* showed that the bacteria express the two enzymes PETase and MHETase, which contribute to PET degradation (3). PETase is the enzyme that enables *I. sakaiensis* to successfully degrade PET (3).

However, PETase from *I. sakaiensis* has a low efficiency in breaking down PET. Factors that affect the rate of PET degradation by PETase, according to Kawai *et al.*, include surface topology of the enzyme, water absorbency of PET, and higher enzyme reaction temperatures (4). Changing these factors can result in faster rates of PET degradation by PETase. For maximum efficiency, the optimal reaction temperature (Topt) should be above 60–65 °C because the polymer chain of the plastic fluctuates at these temperatures. This fluctuation allows water molecules to enter between the chains and weaken them, thus improving the efficiency at which an enzyme can break down PET, as showed by Kawai *et al.* (4). PETase is highly at these temperatures due to its low T_{opt}, which greatly slows the rate of plastic degradation.

Directed evolution is a powerful technique in which iterative mutational analysis is used to alter the function of a biological molecule, such as an enzyme, to fit a need (5). First, directed evolution generates different possible mutations of an enzyme. Second, based on those mutations, corresponding mutants are produced and then evaluated in the lab. The best scoring mutants are then selected based on the user-defined goal such as activity or thermostability. Directed evolution then repeats this process with the top mutants from the previous iteration now acting as the main enzyme.

Performing directed evolution using machine learning on the computer is known as *in silico* directed evolution. For *in silico* directed evolution, instead of producing those mutants in the lab, machine learning is used to evaluate different possible mutations of enzymes. Based on the machine learning evaluation, the algorithm then selects the best mutant and uses it as a starting point again. Machine learning-guided directed evolution is beneficial because machine learning algorithms can take in more data at once, iterations are faster, and the process is cheaper and less time consuming than actually performing directed evolution in the lab. One challenge that exists with this method is that if the machine learning algorithms have high performance on the training set (as measured by accuracy or R^2 values), but do not generalize well to real-world data, directed evolution will not achieve its purpose. Another challenge in machine learning-guided directed evolution is using machine learning models to rank enzymes by continuous output variables like T_{opt}. Machine learning models are often accurate at classification tasks, such as predicting whether an enzyme's T_{opt} falls in some range. However, these models are less accurate at predicting continuous output variables, such as what exactly an enzyme's T_{opt} value will be. One case where machine learning-guided directed evolution has been used successfully for enzyme engineering is in 2019, when a group of researchers engineered a new enzyme for stereodivergent carbon–silicon bond formation, a new-to-nature chemical transformation (6). However, machine learning-guided directed evolution has not previously been used to engineer enzymes that break down non-biodegradable plastic.

We hypothesized that machine learning can be used to predict the optimal temperature at which an enzyme will function, and that machine learning can be combined with directed evolution to engineer an optimized mutant of PETase with a T_{opt} greater than 60 °C for more efficient breakdown of PET and nonbiodegradable plastic. In this study, we used machine learning to perform *in silico* directed evolution on the PETase enzyme to design a mutant that has a predicted T_{opt} of 70 °C, nearly double that of the wild type PETase. This novel enzyme has the potential to break down non-biodegradable plastic more efficiently and at a faster rate than PETase by functioning at a higher optimal temperature. This enzyme is also predicted by external algorithms to have a higher thermostability than the original PETase enzyme.

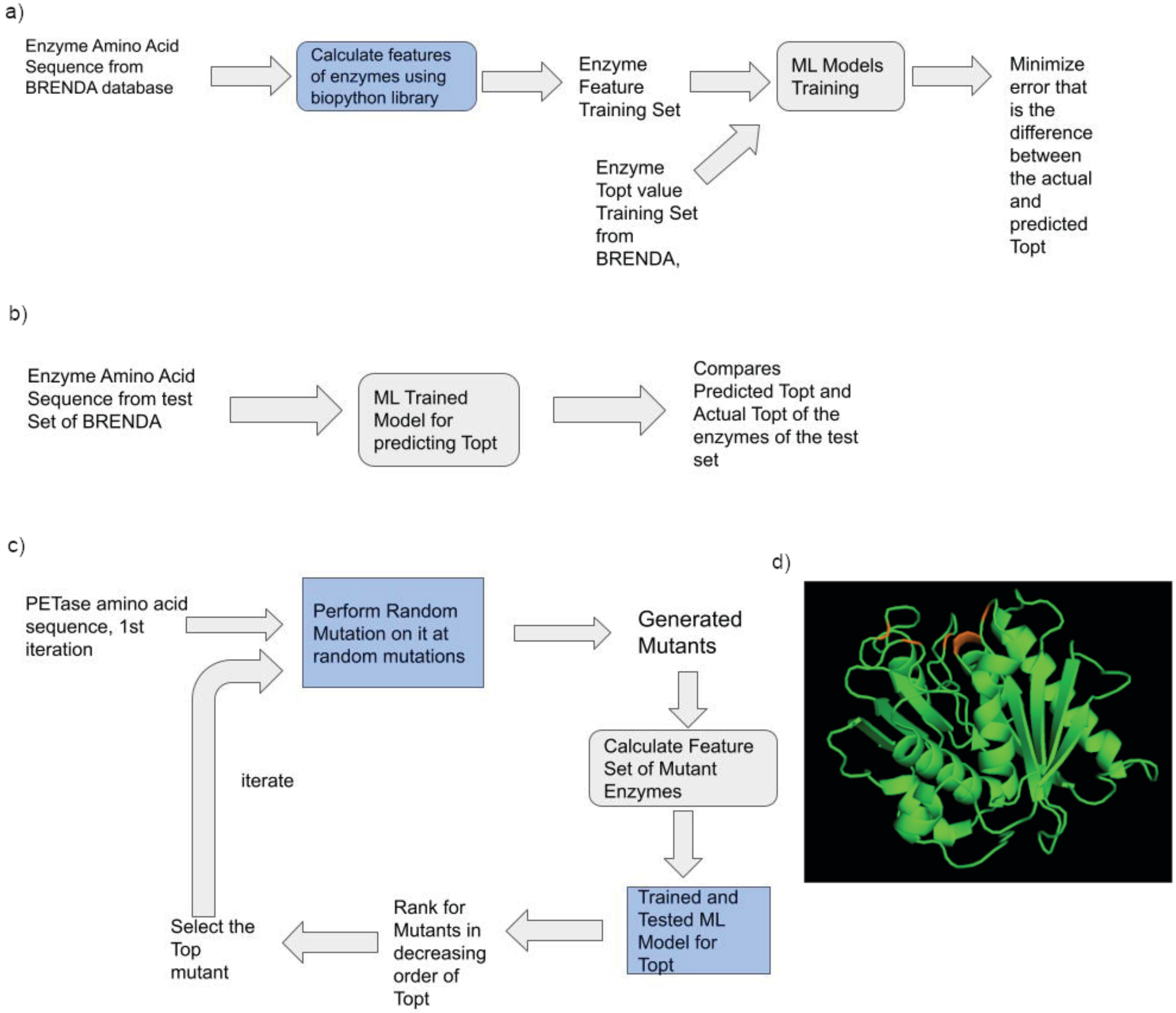

a)
Enzyme Amino Acid Sequence from BRENDA database
Calculate features of enzymes using biopython library
Enzyme Feature Training Set
ML Models Training
Minimize error that is the difference between the actual and predicted Topt
Enzyme Topt value Training Set from BRENDA,
b)
Enzyme Amino Acid Sequence from test Set of BRENDA
ML Trained Model for predicting Topt
Compares Predicted Topt and Actual Topt of the enzymes of the test set
c)
PETase amino acid sequence, 1st iteration
Perform Random Mutation on it at random mutations
Generated Mutants
Calculate Feature Set of Mutant Enzymes
Trained and Tested ML Model for Topt
Rank for Mutants in decreasing order of Topt
Select the Top mutant
iterate
d)

Figure 1: Schematic of machine learning process and PETase structure. a) The training procedure for the Linear Regression, Logistic Regression, and Random Forest Regression models for predicting the T_{opt} on the training set of the data. **b)** The three machine learning models were used after training to predict the T_{opt} of the enzymes in the test set of the data. **c)** The process of performing machine-learning guided *in silico* directed evolution on PETase to improve the enzyme's optimal temperature. **d)** Crystal structure of PETase enzyme from *Ideonella sakaiensis* from the Protein Data Bank (PDB) entry 6ANE with substrate binding residues highlighted in orange (17).

This approach is novel because it is the first to optimize the PETase enzyme optimal reaction temperature using a machine-learning guided directed evolution approach.

First, we used enzyme sequences from the BRENDA database (7) to train two regression models (Linear Regression and Random Forest Regression) to predict the enzyme T_{opt} as well as one classification model (Logistic Regression) to predict whether the enzyme T_{opt} was in the desired range. Then we performed *in silico* directed evolution on PETase by generating mutant enzymes and then screening them against the machine learning models to predict the new T_{opt} of the mutant enzyme and select new mutants for further evolution. We determined that 29 iterations of directed evolution were sufficient to raise the T_{opt} of the enzyme above 60 °C to allow more efficient breakdown of PET, while keeping the active site of the enzyme constant so as not to impact the binding of the enzyme to PETase. A lower number of iterations of directed evolution also prevents the 3D structure of the enzyme from changing significantly as compared to PETase.

RESULTS

We devised a machine learning-guided directed evolution algorithm written in Python to engineer PETase for higher optimal temperature **(Figure 1)**. In order to guide the directed evolution, a machine learning model was written to predict an enzyme's optimal reaction temperature (T_{opt}). Three machine learning models, Random Forest, Linear Regression and Logistic Regression, were trained for this task using enzymes from the BRENDA database. In the second stage, our algorithm generated millions of PETase mutants by randomly mutating different positions of the amino acid sequence. The mutants were scored using the Random Forest Regression algorithm – the machine learning model that performed the best based on the correlation between predicted and actual T_{opt} – to determine which mutation would lead to the highest T_{opt}. The algorithm then reenacted this mutation and selection process with the best scoring mutants, which now acted as the starting point for the next round of directed evolution. To avoid inhibiting the enzyme's function, the residues making up the enzyme's binding site for PET were not mutated.

Machine Learning Models for Predicting Optimal Reaction Temperature

The data set for the machine learning models consisted of enzymes from 11,420 organisms in total obtained from the BRENDA Database (7). Enzymes without T_{opt} values were removed from the dataset. Before performing a cleanup of the data set, there were 2,745 enzyme amino acid sequences comprising the dataset. During the cleanup process, we dropped duplicates, amino acid sequences with a length less than or equal to 7, and enzymes with a T_{opt} equal to or lower than 0 °C. The final training data consisted of 2,643 enzyme amino acid sequences listed with the experimental T_{opt} of each enzyme. The inputs for the models are the enzyme features such as molecular weight, amino acid frequencies, dipeptide frequencies, and the enzyme's host organism's optimal growth temperature **(Table 1)** (14). These enzyme features were calculated from the amino acid sequence listed from the BRENDA database. The data was then split into a training and test set.

The Random Forest and Linear Regression models were evaluated based on the R^2 value between the predicted T_{opt} values and the actual T_{opt} values on the training and test set. A higher R^2 value implies that the correlation between the predicted and actual T_{opt} is greater, indicating a better performing model. On the other hand, the Logistic Regression

Input Feature	Description	Number of Features
Global Protein Features	Included Aromaticity, Hydrophobicity, Instability Index, Charge, Length, Molecular Weight, Aliphaticity, Charge Density, Boman Index, and PI.	10
Amino Acid Frequencies	The occurrence of a particular amino acid in the sequence divided by the length of the sequence.	20
Dipeptide Frequencies	The occurrence of peptides that yields two molecules of amino acid on hydrolysis.	400
Optimal Growth Temperature	The temperature at which the host organism exhibits maximum growth and reproduction.	1
	Total Number of Features	**431**

Table 1: Enzyme features for machine learning models. Global protein features include features based on the entire amino acid sequence of the protein. All these features were used initially to predict enzyme optimal temperature. Feature selection and ranking were later applied to reduce the number of features used in the machine learning algorithms.

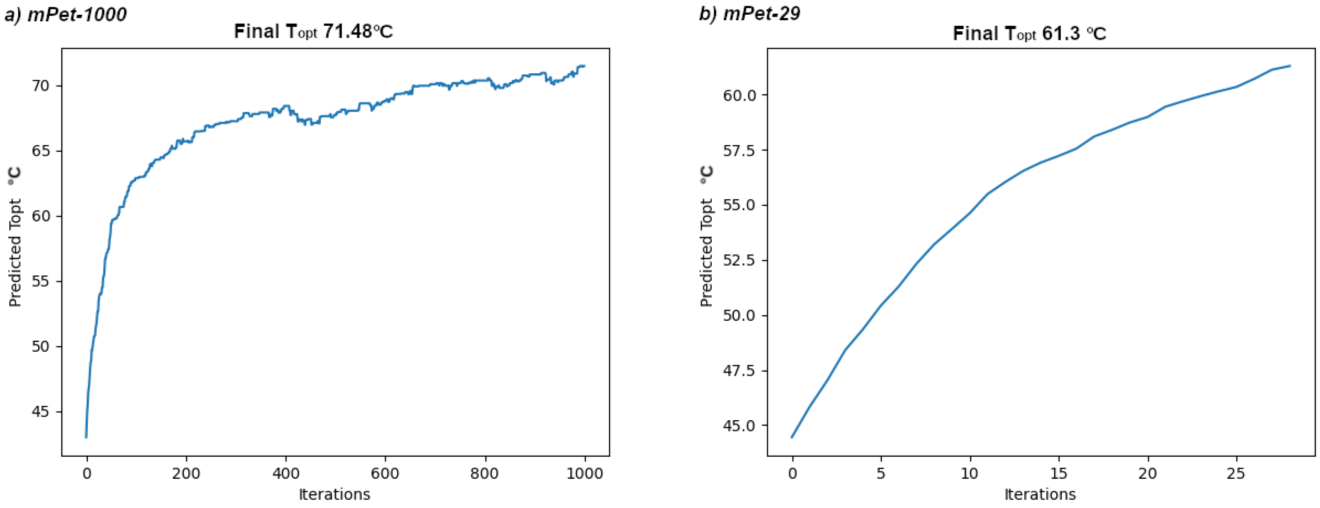


Figure 2: Training and test performance of Lasso Linear Regression and Random Forest Regression models. a) Lasso Linear Regression results on training set, **b)** Lasso Linear Regression results on test set, **c)** Random Forest Regression results on training set, **d)** Random Forest Regression results on test set. The correlation between the model's predicted enzyme T_{opt} values and actual T_{opt} values is shown through the R^2 coefficient. The red line shows y=x, which is interpreted as the ideal situation where there is zero deviation between the model's predicted T_{opt} values and the actual T_{opt} values.

model was evaluated using an accuracy score, which was calculated by the percentage of data points for which the model correctly predicted T_{opt} >=65 °C. Linear Regression obtained an R^2 value of 0.54 on the training set and 0.52 on the test set. Random Forest attained a R^2 value of 0.9322 on the training set and 0.624 on the test set **(Figure 2)**. Logistic Regression achieved a classification accuracy of 92.6% on the training set and 88.3% on the test set. The model with the lowest error, in our case Random Forest Regression, was used to perform directed evolution in the next stage of the approach.

Feature Ranking and Selection

Using Lasso Linear Regression, we ranked the enzyme features that we used as inputs for the models and algorithms and isolated the 10 most important features and their coefficients **(Table 2)**. Out of 431 features, only 156 features had non-zero coefficients and were kept by Lasso Linear Regression in the feature set. Non-zero coefficients mean that the model is taking those features into account when making final predictions of Topt. Higher absolute values of the coefficients correspond to greater feature importance in the final model and higher ranking during the feature selection process. The most important feature by far was the optimal growth temperature (OGT) of the host organism of the enzyme.

Directed Evolution for PETase Engineering

The amino acid sequence for the original PETase enzyme has a Topt of 42 °C **(Table 3)**. The PETase mutant generated after 1000 mutations had a predicted T_{opt} of 71.38 °C **(Table 3, Figure 3a)**. The PETase mutant generated after 29 mutations had a predicted Topt of 61.3 °C **(Table 3, Figure 3b)**. The

Feature Name	Coefficient
OGT	11.958
Dipeptide Freq: YG	1.044
Dipeptide Freq: YN	0.831
Dipeptide Freq: WG	0.733
Amino Acid H Frequency	-0.730
Dipeptide Freq: YL	-0.701
Dipeptide Freq: IG	-0.665
Amino Acid Y Frequency	0.616
Dipeptide Freq: PG	0.588
Dipeptide Freq: YA	0.558

Table 2: After conducting Lasso Linear Regression, the coefficient for each feature was extracted from the model. Only 156 out of 431 features had non-zero coefficients. The features were ranked by importance in decreasing order of the absolute value of their coefficients. A negative sign on the coefficient shows that feature was inversely correlated with predicted optimal temperature. A positive sign indicates that feature was positively correlated with predicted optimal temperature in the model. Features with positive values are highlighted in green and features with negative values are highlighted in red. The top 10 features are presented here.

Type of enzyme	Original PETase enzyme	Modified new PETase enzyme after 1000 iterations (mPet-1000)	Mutant PETase Enzyme after 29 iterations (mPet-29)
Predicted T_{opt}	43.35℃	71.48℃	61.3℃
Melting Temperature	55-65℃	55-65℃	55-65℃
TM Index	0.7715	0.458	**0.988**

Table 3: Comparison between original PETase enzyme and newly designed mutant enzymes after 1000 iterations of directed evolution (mPet-1000) and 29 iterations (mPet-29). For each enzyme, the predicted enzyme optimal temperature (T_{opt}) is provided from the Random Forest Regression model trained in this project. The melting temperature (T_m) range and TM index are also shown, as calculated by the formula published by Ku *et al.* (8).

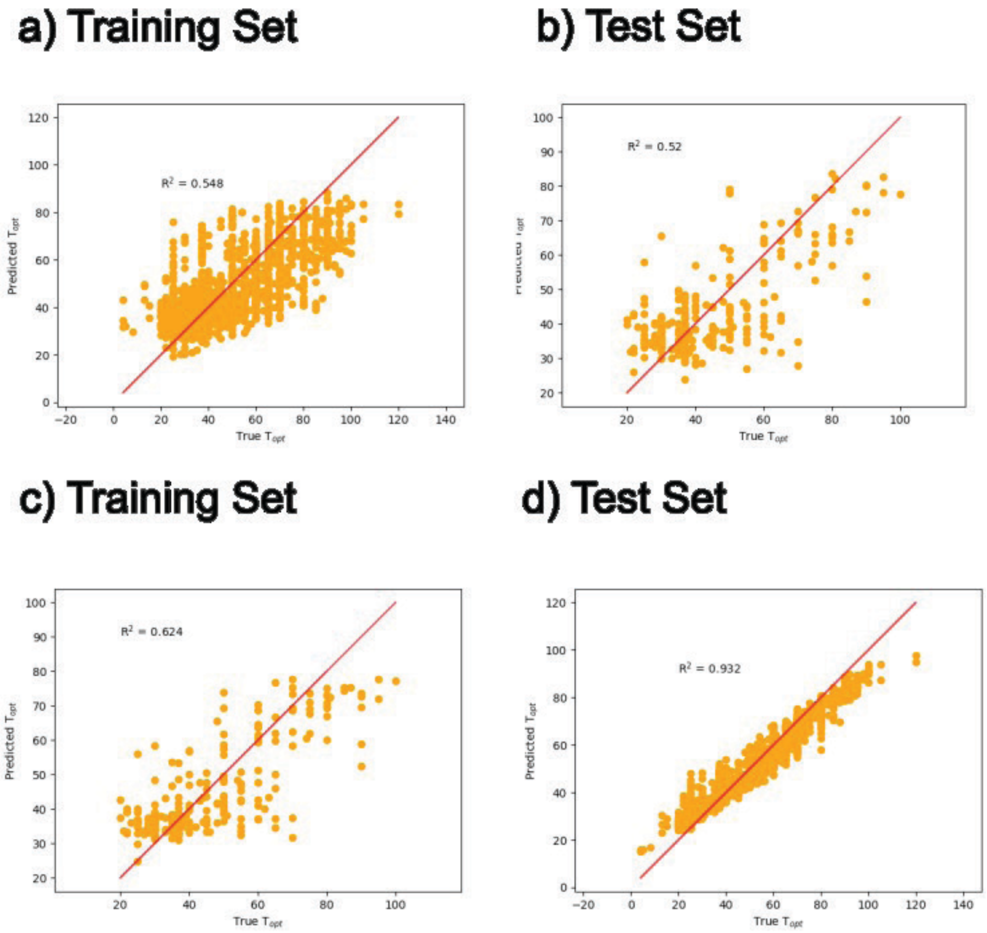
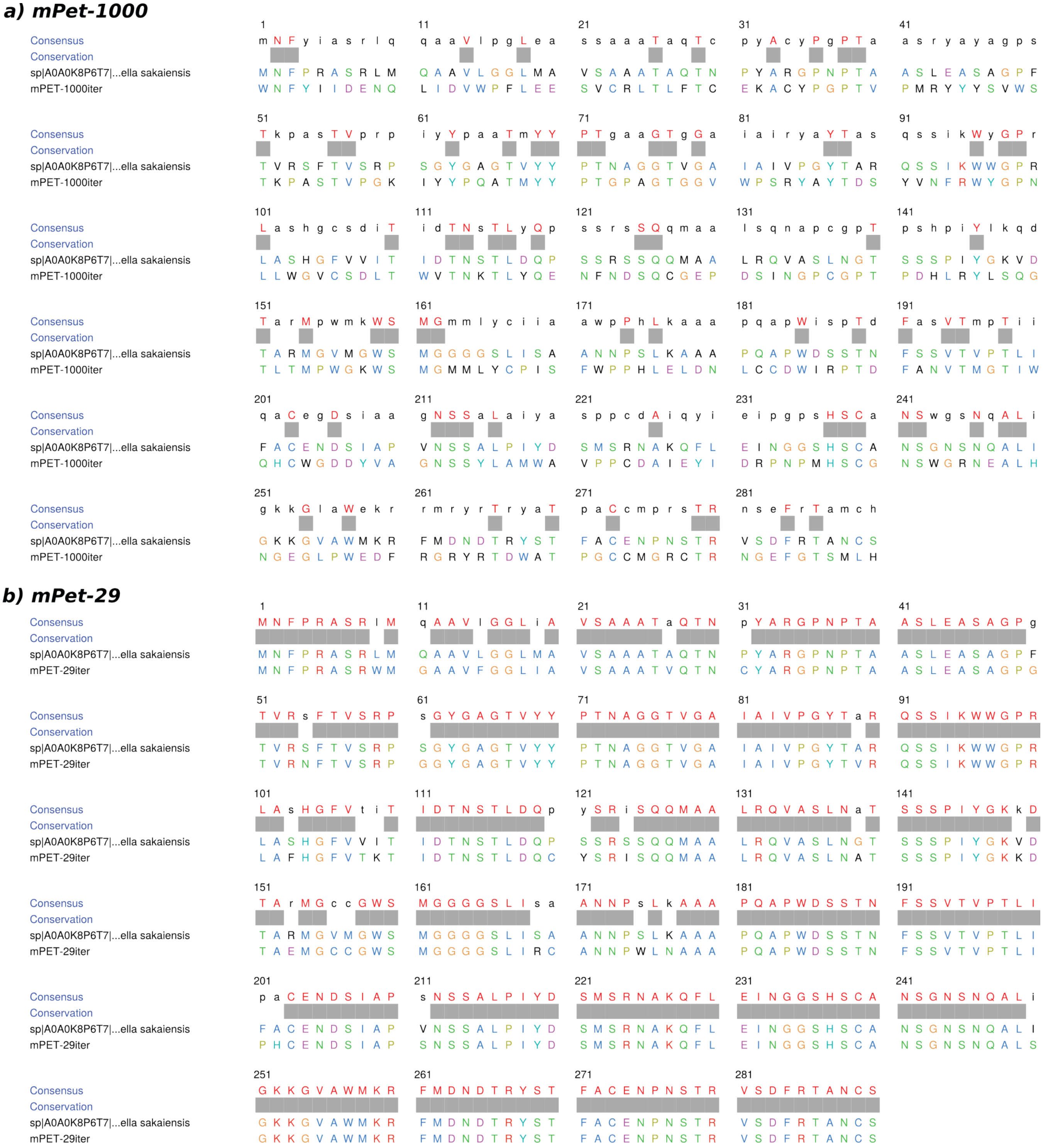


Figure 3: The change in predicted enzyme T_{opt} over 1000 and 29 iterations of directed evolution. a) 1000 iterations of mutations and machine learning-guided directed evolution on the original PETase enzyme. The final Topt we obtained was 71.48 °C. b) 29 iterations of mutations and machine learning-guided directed evolution on the original PETase enzyme. The final Topt we obtained was 60.783 °C.



a) mPet-1000

```
                                   1                    11                   21                   31                   41
Consensus                          m N F y i a s r l q  q a a V l p g L e a  s s a a a T a q T c  p y A c y P g P T a  a s r y a y a g p s
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   M N F P R A S R L M  Q A A V L G G L M A  V S A A A T A Q T N  P Y A R G P N P T A  A S L E A S A G P F
mPET-1000iter                      W N F Y I I D E N Q  L I D V W P F L E E  S V C R L T L F T C  E K A C Y P G P T V  P M R Y Y Y S V W S

                                   51                   61                   71                   81                   91
Consensus                          T k p a s T V p r p  i y Y p a a T m Y Y  P T g a a G T g G a  i a i r y a Y T a s  q s s i k W y G P r
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   T V R S F T V S R P  S G Y G A G T V Y Y  P T N A G G T V G A  I A I V P G Y T A R  Q S S I K W W G P R
mPET-1000iter                      T K P A S T V P G K  I Y Y P Q A T M Y Y  P T G P A G T G G V  W P S R Y A Y T D S  Y V N F R W Y G P N

                                   101                  111                  121                  131                  141
Consensus                          L a s h g c s d i T  i d T N s T L y Q p  s s r s S Q q m a a  l s q n a p c g p T  p s h p i Y l k q d
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   L A S H G F V V I T  I D T N S T L D Q P  S S R S S Q Q M A A  L R Q V A S L N G T  S S S P I Y G K V D
mPET-1000iter                      L L W G V C S D L T  W V T N K T L Y Q E  N F N D S Q C G E P  D S I N G P C G P T  P D H L R Y L S Q G

                                   151                  161                  171                  181                  191
Consensus                          T a r M p w m k W S  M G m m l y c i i a  a w p P h L k a a a  p q a p W i s p T d  F a s V T m p T i i
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   T A R M G V M G W S  M G G G G S L I S A  A N N P S L K A A A  P Q A P W D S S T N  F S S V T V P T L I
mPET-1000iter                      T L T M P W G K W S  M G M M L Y C P I S  F W P P H L E L D N  L C C D W I R P T D  F A N V T M G T I W

                                   201                  211                  221                  231                  241
Consensus                          q a C e g D s i a a  g N S S a L a i y a  s p p c d A i q y i  e i p g p s H S C a  N S w g s N q A L i
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   F A C E N D S I A P  V N S S A L P I Y D  S M S R N A K Q F L  E I N G G S H S C A  N S G N S N Q A L I
mPET-1000iter                      Q H C W G D D Y V A  G N S S Y L A M W A  V P P C D A I E Y I  D R P N P M H S C G  N S W G R N E A L H

                                   251                  261                  271                  281
Consensus                          g k k G l a W e k r  r m r y r T r y a T  p a C c m p r s T R  n s e F r T a m c h
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   G K K G V A W M K R  F M D N D T R Y S T  F A C E N P N S T R  V S D F R T A N C S
mPET-1000iter                      N G E G L P W E D F  R G R Y R T D W A T  P G C C M G R C T R  N G E F G T S M L H
```

b) mPet-29

```
                                   1                    11                   21                   31                   41
Consensus                          M N F P R A S R l M  q A A V l G G L i A  V S A A A T a Q T N  p Y A R G P N P T A  A S L E A S A G P g
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   M N F P R A S R L M  Q A A V L G G L M A  V S A A A T A Q T N  P Y A R G P N P T A  A S L E A S A G P F
mPET-29iter                        M N F P R A S R W M  G A A V F G G L I A  V S A A A T V Q T N  C Y A R G P N P T A  A S L E A S A G P G

                                   51                   61                   71                   81                   91
Consensus                          T V R s F T V S R P  s G Y G A G T V Y Y  P T N A G G T V G A  I A I V P G Y T a R  Q S S I K W W G P R
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   T V R S F T V S R P  S G Y G A G T V Y Y  P T N A G G T V G A  I A I V P G Y T A R  Q S S I K W W G P R
mPET-29iter                        T V R N F T V S R P  G G Y G A G T V Y Y  P T N A G G T V G A  I A I V P G Y T V R  Q S S I K W W G P R

                                   101                  111                  121                  131                  141
Consensus                          L A s H G F V t i T  I D T N S T L D Q p  y S R i S Q Q M A A  L R Q V A S L N a T  S S S P I Y G K k D
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   L A S H G F V V I T  I D T N S T L D Q P  S S R S S Q Q M A A  L R Q V A S L N G T  S S S P I Y G K V D
mPET-29iter                        L A F H G F V T K T  I D T N S T L D Q C  Y S R I S Q Q M A A  L R Q V A S L N A T  S S S P I Y G K K D

                                   151                  161                  171                  181                  191
Consensus                          T A r M G c c G W S  M G G G G S L I s a  A N N P s L k A A A  P Q A P W D S S T N  F S S V T V P T L I
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   T A R M G V M G W S  M G G G G S L I S A  A N N P S L K A A A  P Q A P W D S S T N  F S S V T V P T L I
mPET-29iter                        T A E M G C C G W S  M G G G G S L I R C  A N N P W L N A A A  P Q A P W D S S T N  F S S V T V P T L I

                                   201                  211                  221                  231                  241
Consensus                          p a C E N D S I A P  s N S S A L P I Y D  S M S R N A K Q F L  E I N G G S H S C A  N S G N S N Q A L i
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   F A C E N D S I A P  V N S S A L P I Y D  S M S R N A K Q F L  E I N G G S H S C A  N S G N S N Q A L I
mPET-29iter                        P H C E N D S I A P  S N S S A L P I Y D  S M S R N A K Q F L  E I N G G S H S C A  N S G N S N Q A L S

                                   251                  261                  271                  281
Consensus                          G K K G V A W M K R  F M D N D T R Y S T  F A C E N P N S T R  V S D F R T A N C S
Conservation
sp|A0A0K8P6T7|...ella sakaiensis   G K K G V A W M K R  F M D N D T R Y S T  F A C E N P N S T R  V S D F R T A N C S
mPET-29iter                        G K K G V A W M K R  F M D N D T R Y S T  F A C E N P N S T R  V S D F R T A N C S
```

Figure 4: A comparison between the amino acids sequence of PETase and those of mPET-29 and mPET-1000. a) Pairwise alignment between amino acid sequences of original PETase (spA0A0K8P6T7|*Ideonella sakaiensis*) and mPET-1000. **b)** Pairwise alignment between original PETase (spA0A0K8P6T7|*I sakaiensis*) and mPET-29. mPet-1000 = Modified new PETase enzyme after 1000 iterations of machine learning-guided directed evolution. mPet-29 = Modified new PETase enzyme after 29 iterations of machine learning-guided directed evolution. Pairwise alignments were conducted using EMBOSS Needle and graphs were made using Chimera. If the amino acid of the mutant sequence is the same as the amino acid of the original sequence at that position, then the consensus sequence shows that amino acid in red. If the two amino acids are different, the consensus sequence shows a randomly selected amino acid at that position in lowercase black. The conservation row is gray in areas where the mutant and original amino acid sequences are the same, and absent where the two sequences are different.

optimal temperature of the mutant enzyme increased steadily through the 29 iterations, however with 1000 iterations the optimal temperature did not increase constantly and instead had a long period of fluctuation **(Figure 3)**.

The differences between the original PETase sequences and engineered enzymes were visualized using pairwise alignments **(Figure 4)**. Two pairwise alignments were conducted: one between the original PETase and the mutant PETase after 1000 iterations of machine learning-guided directed evolution (mPET-1000), and the other between original PETase and the mutant PETase after 29 iterations (mPET-29). Pairwise alignments show that mPET-1000 is almost completely a new enzyme with little similarity to the original enzyme. However, mPET-29 is much more similar in its sequence to the original PETase. mPET-29 has 29 mutations out of 290 amino acids, so no more than 10% of the original amino acids are mutated and none of the active site residues are mutated. mPET-29 also has better predicted thermostability than mPET-1000. Consequently, mPET-29 is a promising enzyme for future studies.

The thermostability of the enzymes were measured by the predicted melting temperatures **(Table 3)**. We used the melting temperature predictor and TM index published by Ku *et al.* to estimate the melting temperatures of the enzyme from its sequence (8). The TM index is a way to characterize the thermostability of the amino acid sequence that correlates to the predicted melting temperature of the sequence. The TM index was found to be proportional to the actual melting temperatures of the enzymes. TM Index > 1 implies that the true melting temperature of the protein is likely to exceed 65 °C, whereas a TM < 0 implies that the true melting temperature is likely to be below 55 °C. A TM index between 0 and 1 implies that the true melting temperature is within 55 and 65 °C, with higher TM indexes correlated with higher melting temperatures.

The original PETase enzyme had a melting temperature in the range of 55 to 65 °C and a TM index of 0.778. mPET-1000 was also predicted as having a melting temperature in the range of 55 to 65 °C and achieved a TM index of 0.458. mPET-29 was also predicted as having a melting temperature in the range of 55 to 65 °C and received a TM index of 0.988 **(Table 3)**. These melting temperatures further validate that the mutant enzymes produced by the algorithm are stable and can function at their optimal temperatures without degrading. These values show that the PETase and the mutant PETase enzymes' thermostability is between 55 and 65 °C. The TM index for the mutated PETase enzyme is higher than that of the original PETase enzyme, signifying that increasing the enzyme's Topt also increased the enzyme's thermostability.

DISCUSSION

Out of the three machine learning models trained, Random Forest was the best regression model for calculating the actual Topt, while Logistic Regression was the best classifier for predicting Topt above 65 °C. Lasso Linear Regression was also used to rank the input features by their coefficients when predicting T_{opt}. The ranking validated that the OGT of the enzyme's host organism was the most important factor in predicting T_{opt}. This makes sense because the T_{opt} would naturally evolve to be around the range of the temperature where its host organism grows. For example, enzymes of thermophilic organisms would have to have a high T_{opt} in order for the enzyme to function at the high temperatures these organisms thrive in.

As seen by the R^2 value of 0.54 on the training set and 0.52 on the test set, Linear Regression with Lasso Regularization did not work very well on both of the datasets for predicting the T_{opt} of the enzymes. However as shown by its similar R^2 values on both the training and test set, Linear Regression generalized well. With an R^2 value of 0.9322 on the training set and 0.624, Random Forest did very well on the training set and was the best regression model for predicting T_{opt}. Logistic Regression showed a high accuracy of 92.6% on the training set and of 88.3% on the test set for predicting a Topt greater than or equal to 65 °C. As compared to a previous research study that created a model to predict T_{opt}, our algorithm performed better based on the performance metrics of the R^2 score (9). The other model achieved a R^2 score of 0.51 on the test set using Random Forest and Deep Learning. We believe that our algorithm likely outperformed the existing model due to the inclusion of additional newly released training data and the inclusion of feature selection to reduce overfitting.

In the directed evolution stage of the algorithm, mPET-29 is the better enzyme to test in the lab as it more closely resembles the original PETase enzyme as compared to mPET-1000. mPET-1000 achieved a T_{opt} of 71.38 °C whereas mPET-29 achieved a T_{opt} of 61.3 °C. The T_{opt} of the enzyme increased at a very high rate for the first 200 iterations of Directed Evolution and then stabilized and continued to rise between 65 and 70 °C **(Figure 3b)**.

mPET-1000 could be considered as a separate enzyme rather than a mutant of PETase, since its amino acid sequence differs greatly from that of the original PETase enzyme. In addition, mPET-29 achieved a much higher TM index (0.988) than the mutant PETase enzyme after 1000 iterations (0.458). This might be due to the TM predictor considering mPET-1000 to be separate from the original PETase enzyme since their amino acid sequences vary so greatly. mPET-29 more closely resembles the original PETase enzyme.

These enzyme-melting temperature values validate that the PETase and the mutant PETase enzymes' thermostability is between 55 and 65 °C. This means that we can indeed optimize the optimal temperatures of PETase above 60 °C for maximum efficiency, and the enzyme will still be stable and working at those temperatures. In addition, mPET-29 has a higher TM index as compared to the original PETase enzyme, which provides external validation that our machine learning-guided directed evolution was successful. This development also signifies that increasing the enzyme's T_{opt} also increased its thermostability.

In order to avoid potential sources of error, while generating mutants, we designed the algorithm to avoid mutating the substrate site of the PETase enzyme in order to ensure the enzyme would continue to function. This was implemented by preserving the substrate binding positions on the amino acid sequence and making sure the algorithm excluded those areas while generating mutants.

In the future, we plan to synthesize and test the two mutant enzymes, mPET-29 and mPET-1000, in the lab. We will also test their optimal temperature for breaking down PET. We will express the enzymes into bacteria and then measure the bacteria's efficiency in degrading PET samples of differing sizes. Another interesting approach is to express the best mutant enzyme in cyanobacteria, in order to allow these photosynthetic bacteria (which include blue-green algae) to degrade the plastic in the sea and convert it into non-harmful products for the ocean and marine life. While the mutant enzyme's T_{opt} does not match that of cyanobacteria ocean environments, which are typically 27 °C, we hypothesize that these enzymes will still be able to function in cyanobacteria as the enzymes will remain stable there. In addition, by extending the directed evolution approach presented in this project, PETase variants can be created that have optimal temperatures suitable for the oceanic environments that cyanobacteria inhabit.

MATERIALS AND METHODS

Machine Learning Main Procedure and Training the Machine Learning Models

Enzyme data from the Brenda Database was used to train the machine learning models to predict the enzyme's T_{opt}. After obtaining the data, the algorithm split the data set into training (90% of the data set) and independent test sets (10% of the data set). The inputs for the model were the enzyme features calculated from the amino acid sequence **(Table 1)**. These inputs included amino acid frequency, dipeptide frequency, and optimal growth temperature (OGT). Optimal growth temperature is defined as the temperature at which the host organism of the enzyme has maximum growth and reproduction. OGT was listed in the BRENDA database. The remaining features were calculated from the amino acid sequences through the *modLAMP* Python library (12). Before performing a cleanup of the BRENDA data set, there were 2745 enzyme amino acid sequences comprising the dataset. During the cleanup process, we used the *pandas* library to drop duplicate rows, amino acid sequences with a length less than or equal to seven, and enzymes with a T_{opt} equal to or lower than 0 °C. The final training data consisted of 2643 enzyme amino acid sequences listed with the experimental T_{opt} of each enzyme.

Three Machine learning models were trained to predict Topt: Random Forest, Linear Regression and Logistic regression. The three Machine Learning models were implemented using the *scikit-learn* Python library (11).

Classification Models such as Logistic Regression are concerned with predicting a discrete label, such as $T_{opt} \geq 65$ °C. Regression models such as Linear Regression and Random Forest Regression models fundamentally predict a continuous quantity such as the actual T_{opt} value for the enzymes. In particular, Random Forest operates by constructing a multitude of decision trees during training and outputting the mean of prediction of the individual trees. Decision trees are structures in which each internal node represents a possible value of an input feature. In Random Forest models, several decision trees are constructed based on the input training data. At test time, each decision tree produces a real value that is a prediction of the output variable. To obtain the final prediction for the output variable, which is T_{opt} in this case, the predictions from each individual decision tree are averaged.

In the training stage, first Linear Regression was trained to have a high R^2 value (correlation value) between the predicted and the actual T_{opt} of the enzymes. To correct the overfitting resulting from an analysis of linear regression, we implemented lasso regression. Lasso Regression is linear regression with added regularization. This regularization adds a penalty to a model if the model is overfitting, meaning it is setting the value of its weights too high. Here, Lasso Regularization set the weights of the enzyme features that were unimportant to Linear Regression to 0.

Following linear regression, a Random Forest regression model was trained to predict T_{opt}. Random Forest models operate by assembling different decision trees based on the input data, taking the mean of those decision trees and setting that as the predicted output. Logistic Regression was the classifier and the classifier performance was measured by accuracy in predicting whether an enzyme's Topt was above 65 °C. The graphs of performance for Linear Regression and Random Forest Regression were generated using the *matplotlib* Python library.

Feature Selection and Ranking

To minimize overfitting, Lasso Regression includes feature selection, which is the process of only including features with non-zero coefficients as inputs to the model to calculate the output value of the predicted T_{opt} of the enzyme. The absolute values of the coefficients for each feature in a linear regression model determine how much importance that feature is given in predicting the final output. The predicted output value, y_{pred}, was calculated by the following equation, where x_i represents each input feature and a_i represents the coefficient for each input feature:

$$y_{pred} = \sum a_i * x_i$$

The coefficients are iteratively determined to the values that minimize the prediction error on the training dataset. As seen by the previous equation, input features with low coefficient values have less impact on the output variable. Consequently, input features can be ranked by importance in decreasing order of the absolute values of their coefficients. Essentially, Lasso Regression implements feature selection by setting the coefficients of features with low importance to zero, unlike traditional linear regression, which gives those features low, but non-zero. This process implemented in Lasso is known as L1 regularization.

This led us to a feature set which consisted of a Aliphaticity, Charge Density, Boman Index, and the isoelectric point (PI). Aliphaticity refers to the aliphatic nature of the chemical compounds, indicating that the amino acid side chains of the enzymes only contain carbon or hydrogens. Charge Density indicates the distribution of charge across the amino acids. The Boman index indicates the potential binding interactions the enzyme can have, and the index value is equated with the solubility values for all residues in a sequence. PI is the pH value where the net charge of the enzyme is 0.

Directed Evolution Procedure

First, in our machine learning-guided directed evolution approach, the algorithm randomly mutated the PETase enzyme at random positions, excluding the substrate site at positions where the PET molecule binds to the enzyme according to the Uniprot database (13). One thousand mutants were generated in this way. Second, based on those mutations, the algorithm used Random Forest to score the corresponding mutants based on their predicted T_{opt} values. The algorithm then selected the best scoring mutant and performed random mutations on it at random positions, again excluding the substrate site. This directed evolution process was repeated for 1000 iterations. However, 1000 iterations would lead to approximately 1000 mutations in the original enzyme, which would create an enzyme vastly different from the original PETase. The original PETase enzyme has only 290 amino acids. Thus, this process was then also repeated for 29 iterations in order to yield a mutant enzyme more similar to PETase rather than a completely new enzyme. Two pairwise sequence alignments were conducted: one between the amino acid sequences of original PETase and mPET-1000, and the other between original PETase and mPET-29. Pairwise alignments were conducted using the EMBOSS Needle webserver, which uses the Needleman-Wunsch algorithm for global sequence alignment. Visualizations were made using the alignment visualization tool within Chimera (14, 15, 16).

Directed Evolution Procedure

The enzymes' melting temperature ranges and TM index values were calculated by the online melting temperature predictor by Ku *et al.* (8). The TM index is calculated based on the length of the amino acid sequence and the composition of dipeptides within the protein sequence:

$$TI = \frac{(100/L)^{*} \sum P_{index}(x_i y_{i+1}) - 9372}{398}$$

Received: January 19, 2022
Accepted: July 1, 2022
Published: January 31, 2023

REFERENCES

1. "A Million Bottles a Minute: World's Plastic Binge 'as Dangerous as Climate Change'." 28 June 2017, www.theguardian.com/environment/2017/jun/28/a-million-a-minute-worlds-plastic-bottle-binge-as-dangerous-as-climate-change. Accessed 2 Aug. 2022.
2. Parker, Laura. "A Whopping 91 Percent of Plastic Isn't Recycled | National Geographic Society." 5 July 2019, www.nationalgeographic.org/article/whopping-91-percent-plastic-isnt-recycled/. Accessed 2 Aug. 2022.
3. Yoshida, Shosuke, *et al.* "A Bacterium That Degrades and Assimilates Poly(Ethylene Terephthalate)." *Science,* vol. 351, no. 6278, American Association for the Advancement of Science, Mar. 2016, pp. 1196–99. doi:10.1126/science.aad6359.
4. Kawai, Fusako, *et al.* "Current Knowledge on Enzymatic PET Degradation and Its Possible Application to Waste Stream Management and Other Fields." *Applied Microbiology and Biotechnology,* vol. 103, no. 11, June 2019, pp. 4253–68. *Springer Link,* doi: 10.1007/s00253-019-09717-y.
5. Frances H. Arnold - Nobel Lecture: Innovation by Evolution - NobelPrize.Org. https://www.nobelprize.org/prizes/chemistry/2018/arnold/lecture/.
6. Wu, Zachary, *et al.* "Machine learning-assisted directed protein evolution with combinatorial libraries." *Proceedings of the National Academy of Sciences* 116.18 (2019): 8852-8858.
7. Schomburg, Ida *et al.* "BRENDA, enzyme data and metabolic information." *Nucleic Acids Research* vol. 30,1 (2002): 47-9. doi:10.1093/nar/30.1.47
8. Ku, Tienhsiung, *et al.* "Predicting Melting Temperature Directly from Protein Sequences." *Computational Biology and Chemistry,* Elsevier, 20 Oct. 2009, doi:10.1016/j.compbiolchem.2009.10.002
9. Gado, Japheth E., *et al.* "Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning." *Journal of Chemical Information and Modeling,* American Chemical Society, 8 July 2020, doi:10.1021/acs.jcim.0c00489.
10. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996): 267-288 doi:10.1111/j.2517-6161
11. Pedregosa, Fabian, *et al.* "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research,* 1 Jan. 1970, jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
12. Müller, Alex T, *et al.* "ModlAMP: Python for Antimicrobial Peptides." OUP Academic, *Oxford University Press,* 4 May 2017, doi:10.1093/bioinformatics/btx285
13. UniProt Consortium. "UniProt: a hub for protein

information." *Nucleic acids research* 43.D1 (2015): D204-D212. doi:10.1093/nar/gku989

14. Sauer, David B., and Da-Neng Wang. "Predicting the optimal growth temperatures of prokaryotes using only genome derived features." *Bioinformatics* 35.18 (2019): 3224-3231 doi:10.1093/bioinformatics/btz059
15. Madeira, Fábio, *et al.* "The EMBL-EBI search and sequence analysis tools APIs in 2019." *Nucleic Acids Research* 47.W1 (2019): W636-W641. doi:10.1093/nar/gkz268
16. Pettersen, Eric F., *et al.* "UCSF Chimera—a visualization system for exploratory research and analysis." *Journal of Computational Chemistry* 25.13 (2004): 1605-1612. doi:10.1002/jcc.20084
17. Fecker, Tobias, *et al.* "Active site flexibility as a hallmark for efficient PET degradation by I. sakaiensis PETase." *Biophysical Journal* 114.6 (2018): 1302-1312. doi:10.1016/j.bpj.2018.02.005