Current Research in Biotechnology 4 (2022) 1–9

Contents lists available at ScienceDirect

Current Research in Biotechnology

journal homepage: www.elsevier.com/locate/crbiot

EnZymClass: Substrate specificity prediction tool of plant acyl-ACP thioesterases based on ensemble learning

Deepro Banerjee [a,1], Michael A. Jindra [b,1], Alec J. Linot [b], Brian F. Pfleger [b], Costas D. Maranas [c,*]

[a] The Bioinformatics and Genomics Program, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, United States
[b] Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI, United States
[c] Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, United States

ARTICLE INFO

Keywords:
Thioesterase
Enzyme classification
Machine learning
Substrate specificity
Medium-chain oleochemicals
Synthetic biology

ABSTRACT

Characterizing the functional properties of plant acyl-ACP thioesterases (TEs), a key enzyme class used in the production of renewable oleochemicals in microbial hosts, experimentally, can be an expensive and time consuming process since it requires manual screening of thousands of candidates in a database. Using amino acid sequence to computationally predict an enzyme's function might accelerate this process; however obtaining the necessary amount of information on previously characterized enzymes and their respective sequences required by standard Machine Learning (ML) based approaches to accurately infer sequence-function relationships can be prohibitive, especially with a low-throughput testing cycle. Experimental noise, unbalanced dataset where high sequence similarity does not always imply identical functional properties will further prevent robust prediction performance. Herein we present a ML method, **En**semble method for en**Zym**e **Class**ification (EnZymClass), that is specifically designed to address these issues. We used EnZymClass to classify TEs into short, long and mixed free fatty acid substrate specificity categories. While general guidelines for inferring substrate specificity have been proposed before, prediction of chain-length preference from primary sequence has remained elusive for plant acyl-ACP TEs. By applying EnZymClass to a subset of TEs in the ThYme database, we identified two medium chain TEs, ClFatB3 and CwFatB2, with previously uncharacterized activity in *E. coli* fatty acid production hosts.

EnZymClass can be readily applied to other protein classification challenges and is available at: https://github.com/deeprob/ThioesteraseEnzymeSpecificity.

1. Introduction

Machine Learning (ML) models are effective tools for narrowing the vast search space in complex biological problems. These tools are apt in protein engineering and bioprospecting applications, as testing every possible residue substitution, or even every available homolog, is experimentally infeasible (Greenhalgh et al., 2018; Yang et al., 2019). However, the efficacy of a ML model is dependent on the availability of an appropriately sized and balanced training dataset, with the requirement for known inputs scaling with the complexity and number of features needed to describe the data. Thus, a common barrier to utilizing data science for facilitation of experimental efforts is the compilation of a comprehensive dataset to train a predictive and accurate ML model. The development of a predictive model which can be trained with dataset sizes on the order of what is available to experimentalists, namely less than 1,000 training instances, would serve to facilitate a protein discovery and engineering pipeline for development of biocatalysts, biologic pharmaceuticals, and membrane transporters with a desired function. In this study, we apply ML to the discovery of substrate-specific enzymes within the broader family of acyl-acyl carrier protein (ACP) thioesterases (TEs) with a training dataset of less than 120 characterized sequences in the academic and patent literature. We then use the ML model to identify acyl-ACP TEs whose sequences indicate medium-chain specificity. Finally, we complement the results from our search by incorporating activity-enhancing modifications into the sequences of the newly identified TEs.

ML approaches which have been demonstrated to infer protein function and enzyme substrate specificity (Amin et al., 2013; Khurana et al., 2010) from primary sequence fall under two categories:

* Corresponding author.
E-mail address: cdm8@psu.edu (C.D. Maranas).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.crbiot.2021.12.002
Received 7 September 2021; Revised 2 December 2021; Accepted 12 December 2021

generative and discriminative (Leslie et al., 2002). Recent results suggest that discriminative approaches are superior both in accuracy and computational efficiency of solving the protein classification problem (Leslie et al., 2004). Support Vector Machines (SVM) are among the most widely used discriminative learning algorithms for biological sequence classification and have been experimentally proven to achieve up to 10% higher classification accuracy than generative approaches when applied to biologically relevant problems (Deshpande and Karypis, 2002). While effective, SVM classifiers are highly influenced by the feature extraction technique employed to encode the protein sequences (Saigo et al., 2004). Proper selection of a feature extraction method is integral to model performance, especially when available training data is limited.

Feature extraction of protein sequences generates a discrete numerical representation of a protein to create feature vectors that are correlated with the attribute of the protein which will be predicted by the model. To train an SVM, several feature extraction techniques for protein sequences have been implemented. These techniques include kernel-based methods (Leslie et al., 2002), physicochemical encoding of protein sequences (Ding and Dubchak, 2001), N-gram representations (Jurafsky and Martin, 2016), and Position-Specific Scoring Matrices (Nanni et al., 2013). However selection of a specific feature extraction technique may result in a model which loses generalizability across the entire protein classification domain. Recognizing this limitation, Nanni et al. suggested the use of an ensemble of classifiers to attain consistently superior performance over individual feature extraction techniques across the entire protein classification domain (Nanni et al., 2014). This strategy has proven effective in multiple protein classification studies (Çamoğlu et al., 2005; Whalen and Pandey, 2013).

We set out to demonstrate the utility of an ensemble of discriminative ML classification algorithms to broadly predict the presence or absence of a useful protein characteristic from primary sequence information. We selected the detection of medium-chain-length specificity of acyl-ACP TEs as our classification problem due to 1) the relatively small number of experimentally characterized medium chain acyl-ACP TE sequences, 2) the importance of TEs in medium-chain fatty acid synthesis, and 3) the absence of any existing computational tool for this task.

Medium-chain oleochemicals, defined as eight to twelve-carbon free fatty acids and derivatives, are target molecules for synthetic biologists due to limited or challenged accessibility from conventional agricultural or petrochemical routes (S. Kim et al., 2015; Lennen and Pfleger, 2012; Rigouin et al., 2018; Sarria et al., 2017). While these chain lengths have traditionally been sourced from tropical crops, the eight, ten, and twelve-carbon products are not major constituents of the oil (Rupilius and Ahmad, 2007). Processes have been established to create higher value oleochemical derivatives, such as fatty alcohols, directly from petrochemical building blocks. However, these processes yield a distribution of alcohols, and thus do not provide a highly selective route to the medium-chain products (Noweck and Ridder, 1988).

Expression of medium-chain specific acyl-ACP TE is the primary strategy for enabling development of biological platforms to high-value oleochemical derivatives. This approach has facilitated enrichment of product distributions for medium-chain chemical species in bacteria, yeast, and plant systems (Xu et al., 2016). The most specific fatty acid and fatty alcohol distributions have been demonstrated in *E. coli*, with over 90% of product belonging to the C_8 species (Hernández Lozada et al., 2018, 2020). Indeed, the expression of various acyl-ACP TEs, either homologs from nature or variants thereof, has enabled control over the chain-length distribution in *E. coli* (Grisewood et al., 2017; Hernández Lozada et al., 2018) (Fig. 1). Of these studies, acyl-ACP TEs from select plant species have been shown to have greater native specificity toward the medium-chain substrates when compared to bacterial homologs (Sarria et al., 2018). While medium-chain-producing plants are well-catalogued (Ohlrogge et al., 2018), the bioprospecting process to identify novel TEs requires identification of an organism overlooked in the patent literature as well as the isolation of the desired gene from a cDNA library. Thus, the identification of the gene responsible for the medium-chain activity requires further characterization such as *in vivo* production experiments in a genetically tractable host or *in vitro* activity assays (H.J. Kim et al., 2015). This combination of the prolonged testing cycle, the relative rarity of the medium-chain TE, and the high value of medium-chain oleochemical derivatives has resulted in a broadly protected intellectual property landscape (Gordon Roessler and Roy, 2015; Pfleger et al., 2020; Pfleger and Lennen, 2013). Therefore, discovery of novel acyl-ACP TEs is desirable to both navigate the current intellectual property landscape in oleochemical biosynthesis and to elucidate the mechanisms of substrate specificity among this enzyme class.

To bridge the sequence-to-function mapping gap, the ThYme database was compiled to organize putative and characterized thioesterase-active enzymes into various families based on their primary sequence information (Cantu et al., 2010). As a follow-up study, Jing et al. characterized the *in vivo* activity of 24 TEs from the TE14 family in the ThYme database and showed that phylogenetic and sequence identity analysis alone were not sufficient to distinguish plant TEs substrate specificity (Jing et al., 2011).

Herein we put forth a ML-based discriminatory approach, termed EnZymClass (**En**semble method en**Zym**e for **Class**ification), to predict substrate specificity from primary sequence encoded features for TEs uncharacterized in an *E. coli* system. We applied EnZymClass to identify two medium-chain acyl-ACP TEs among a set of 617 TEs catalogued in the ThYme database. We further modified the discovered gene sequences to improve titer and medium-chain specificity over wildtype (WT) when expressed in an *E. coli* production host. This study provides an exemplar of how even limited datasets can be leveraged with ML to support bioprospecting efforts and to provide a suitable starting template enzyme for protein engineering efforts. EnZymClass can be accessed at: https://github.com/deeprob/ThioesteraseEnzymeSpecificity.

2. Materials and methods

2.1. Dataset compilation

2.1.1. Training dataset

The training dataset included primary sequences and *in vivo E. coli* product distributions for 113 acyl-ACP plant TEs previously recorded in various reports from scientific and patent literature (Table S10). Two additional TEs from *Auxenochlorella protothecoides* (KFM28838.1) and *Prunus sibirica L.* (AIX97815.1) were previously tested for free fatty acid production prior to this study and were used to supplement the training dataset. Unique acyl-ACP TE transcripts have been detected in various plant tissue extracts within the same organism, encoding for TE enzymes with different substrate specificities (Dehesh et al., 1996). Furthermore, mutagenesis studies which probed the mechanisms of substrate specificity in the acyl-ACP TE from *Cuphea viscosissima* and *Umbellularia californica* showed that as few as three point mutations in the binding pocket and the ACP binding site can shift observed product distributions in *E. coli* (Jing et al., 2018; Yuan et al., 1995). Therefore, TE homologs from the same organism and variants of CvFatB2 (AEM72523.1) and UcFatB1 (AAA34215.1) were represented in the training dataset. The characterized dataset of 115 plant acyl-ACP TEs is available at https://github.com/deeprob/ThioesteraseEnzymeSpecificity/tree/master/data.
Although many bacterial TE sequences have been previously characterized (Jing et al., 2011), the primary reason we restricted the model to plant TE sequences is the low sequence identity across TEs from the plant and bacterial kingdoms. Bacterial acyl-ACP TE have lower sequence identity (<30%) to the plant acyl-ACP TE, while the

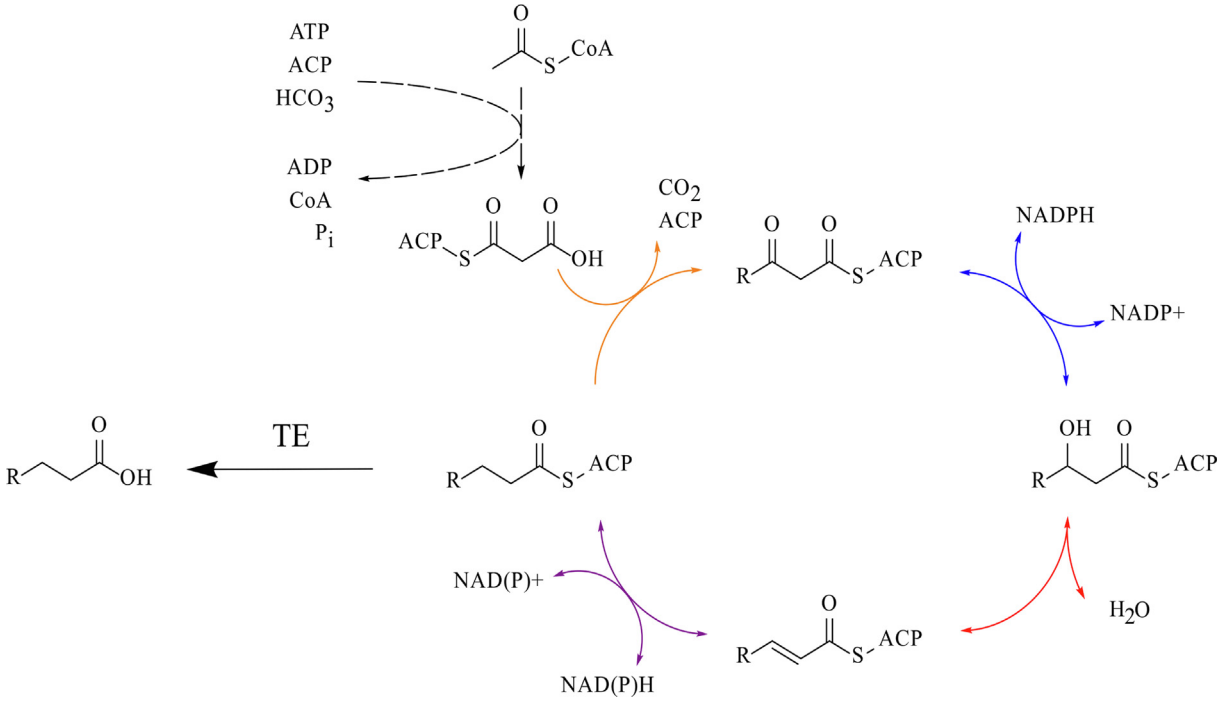

Fig. 1. The acyl-ACP TE plays a key role in fatty acid biosynthesis in *E. coli*. By intercepting the growing acyl-ACP chains, the TE hydrolyzes the acyl chain from the ACP and redirects flux to the free fatty acid pool. These free fatty acids can be further derivatized *in vivo* or *ex vivo*.

sequence identity among the plant TE homologs ranged from ~45% to ~85%. Due to the higher sequence identity among plant sequences, we reasoned the model would make more discerning predictions from whichever amino acid differences may be present among homologs. We were concerned that the model would not accurately distinguish between amino acid changes that were divergent due to evolutionary differences or functional differences if bacterial TE were included in the training dataset. Furthermore, most medium-chain specific acyl-ACP TEs do come from the plant kingdom (Jones et al., 1995). Microbial TEs typically have broad specificity, exhibiting activity on both medium and long-chain substrates, but not showing a strong preference unless further engineered. Each TE included in the training set is presumed to possess the characteristic Hot Dog fold due to the sufficiently high (>25%) sequence identity among all 115 members (Sander and Schneider, 1991). *E. coli* was chosen because it remains the most common and facile method for characterization of heterologous TEs. The product distribution data was subsequently used to classify each TE into three discrete categories: 1) the "medium-chain" category contained TEs which resulted in distributions of at least 50% C_8 to C_{12} free fatty acids, 2) the "long-chain" category contained TEs which produced 50% C_{14} to C_{18} free fatty acids and less than 10% C_8 to C_{12} free fatty acids and, 3) the "mixed distribution" category contained TEs which yielded distributions between 10% and 50% C_8 to C_{12} free fatty acids.

2.1.2. Cross-validation dataset

The dataset of 115 characterized TE sequences was randomly divided into training and validation set using a 75–25 split. This process was simulated 10,000 times yielding different training and validation dataset at each simulation instance. The validation scores reported in this study are the mean scores attained by EnZymClass on 10,000 validation sets.

2.1.3. Test dataset

An independent test set was created by extracting the 617 eukaryotic sequences from the TE14 family in the ThYme database (Cantu et al., 2010). Each sequence in TE14 encodes acyl-ACP TEs from bacterial or plant origins and possesses the characteristic HotDog fold. The test dataset was limited to the sequences denoted with origins from eukaryotic organisms in TE14 to match the training set, which comprised of only plant acyl-ACP TEs.

2.2. Feature extraction

In this work, 47 alignment-free feature extraction techniques that encode primary sequence information of the enzymes into fixed-length feature vectors were employed. The feature extraction techniques fall under four categories, kernel methods, n-gram methods, physicochemical encoding methods, and PSSM profile based methods. TE sequence feature encoding was conducted utilizing source codes from three open-source python or R-based tools, KeBABS (Palme et al., 2015), iFeature (Chen et al., 2018), and POSSUM (Wang et al., 2017). While KeBABS is already an existing R package for kernel methods, we have developed three PyPI (PyPI, 2020) packages (ifeatpro, ngrampro, and pssmpro), for numerical encoding of protein sequences. A brief description of the 47 feature extraction techniques divided into their respective categories along with usage and accessibility details of the protein sequence encoding packages is provided in **Text S1**. The feature extraction category, name, software package used to deploy them and literature from which they are adopted are listed in **Table S1**.

2.3. EnZymClass: Ensemble method for enZyme classification

2.3.1. Model description

EnZymClass consists of N base learners (one for each feature extraction technique) which provide their outputs to a meta learner that predicts the functional attribute of proteins. Although all base learners are trained using the same principle, the heterogeneity among them is governed by the different feature extraction techniques used to encode the set of protein sequences into unique numerical representations. To prevent overfitting, the feature vectors generated through the extraction

process were decomposed into a lower dimensional representation using Principal Component Analysis (PCA) and fed as input to a designated base learner in EnZymClass. The base learner trained on the set of encoded protein sequences yields the predicted functional attribute of a given protein sequence as an output. The outputs of the k-best base learners are passed on to the meta learner that uses a majority voting scheme to predict protein functional attribute category. The k-best base learners are selected through cross-validation. EnZymClass pipeline is presented in **Fig. S1**. The entire workflow of the study including model training, validation, and prediction of uncharacterized TE sequences is illustrated in **Fig. S2** and described in **Text S2**.

2.3.2. Base learner algorithms

EnZymClass supports three base learning algorithms, 1) Support Vector Machine (SVM), 2) Neural Network (NN) and 3) Gradient Boosting Trees (GBC). Selection of the appropriate base learner depends on their cross-validation performance.

2.3.3. The meta learner

The meta learner accepts the outputs of all the base learners as an input vector, implements a hard majority voting scheme, and returns the consensus prediction of the label (TE substrate specificity in our case) as an output. The meta learner output is calculated as follows:

$$\hat{y} = mode\{\hat{y_1}, \hat{y_2}, \cdots, \hat{y_k}\}$$

where $\hat{y}$ is the output of the meta learner, $\hat{y_i}$ is the output of the base learner i, k is the number of best performing base learners to consider, and *mode* is a function that selects the most frequent value among a set of values.

2.3.4. Model hyperparameters

EnZymClass has several hyperparameters which can be tuned to improve model performance. It has a hyperparameter k that denotes the number of best performing base learners that the meta learner needs to take into account. Each base learner has a number of hyperparameters depending on the learning algorithm used to train them. The respective learning algorithm-dependent hyperparameters of the base learners are shown in Table S2. The hyperparameters are learnt through a 5-fold cross validation scheme.

2.3.5. Model evaluation metrics

The performance of EnZymClass was measured using four classification metrics: 1) accuracy score, 2) precision score on the medium chain TE class, 3) recall score on the medium chain TE class, and 4) Matthew's correlation coefficient (MCC). Mathematical formulation of the four performance metrics is given in **Text S3**.

2.4. In vivo characterization of plant acyl-ACP TEs

The cloning of acyl-ACP TE homologs and variants and molecular biology materials are described in **Text S5**. Each TE construct was subsequently transformed into RL08*ara*, an *E. coli* MG1655 derivative engineered for free fatty acid accumulation (Lennen et al., 2010). Single colonies of the RL08 transformants were grown overnight at 37 °C and 250 r.p.m. in 5 mL of LB media supplemented with 100 mg/L of carbenicillin. All strains and vectors used in this study are in **Table S7**.

For free fatty acid production trials, shake flasks containing 25 mL of LB supplemented with 4 g/L of glycerol were inoculated with 275 μL of the stationary phase culture. The cultures were then allowed to grow at 30 °C and 250 r.p.m. until they reached an OD of 0.2–0.3. Isopropyl β-D-thiogalactoside (IPTG) was then added to the media to a final concentration of 20 μM to induce transcription of the TE genes. After the cultures incubated for 24 h at 30 °C and 250 r.p.m., 2.5 mL was sampled for derivatization and characterization of the free fatty acid distribution. For analysis of the free fatty acid distribution present in the supernatant, 10 mL of sample were centrifuged at 4500*g* for 20 min, and 2.5 mL of the supernatant was collected. The fatty acid quantification method described by Politz et al. was then followed with slight modifications (Politz et al., 2016). Namely, the internal standard solution of odd-chain free fatty acids was prepared by combining heptanoic, nonanoic, undecanoic, tridecanoic acid to a final concentration of 5 g/L in methanol, and pentadecanoic and heptadecanoic acid were added to a final concentration of 1 g/L. 100 μL of the internal standard was added to each 2.5 mL sample. After obtaining the chromatograph, each even-chain free fatty acid was quantified by normalizing the peak area of its derivatized methyl ester to the adjacent peaks of the odd-chain methyl esters.

3. Results

3.1. EnZymClass plant acyl-ACP TE substrate specificity prediction

We evaluated the performance of EnZymClass on TE substrate specificity prediction task using a rigorous validation scheme to fairly assess model generalizability, robustness, and reproducibility of results. EnZymClass achieved a mean validation accuracy of 0.8 with a standard deviation of 0.06, mean precision and recall scores of 0.87 and 0.89 respectively for the medium-chain TEs, the product of interest, and mean MCC score of 0.68 across 10,000 simulations using different training and validation sets. The precision (on the medium-chain category), recall (on the medium-chain category), accuracy, and MCC score distributions of EnZymClass are shown in Fig. 2.

Among the three categories of TEs, the mixed specificity class which represented only 17% of all the characterized TEs in our dataset (Fig. S3) was the worst predicted with much lower precision and recall scores compared to the medium-chain and long-chain categories (Table S5). Although EnZymClass offers a mechanism to deal with class imbalance by allowing the user to provide label priority weights as base learner input, we opted against assigning such weights to the minority class. Prioritizing correct classification of the mixed specificity class of TEs would have de-emphasized classification of medium-chain TEs, thus impacting its accurate characterization. The base learners in the ensemble were chosen as SVM classifiers after analyzing the performance of SVM against two learning algorithms: GBC and NN. GBC and NN were both outperformed by SVM in terms of accuracy on separate held-out validation sets (**Table S4**). SVM has higher generalizability and can handle high dimensional datasets (Ben-Hur et al., 2008) which may have resulted in its superior performance on our relatively small dataset. Existing efforts to characterize TEs mostly rely on experimental approaches, which have limited throughput depending on the activity characterization method. Furthermore, previous studies have demonstrated that phylogenetic analysis or sequence similarity-based approaches do not always correlate with TE substrate specificity (Jing et al., 2011, 2018). At present, there is no available computational platform that can match EnZymClass in terms of TE substrate specificity characterization.

3.2. Comparison with individual base learners in EnZymClass

The primary purpose of using an ensemble framework in EnZymClass was to decrease variance in model prediction resulting from a small and unbalanced training set. Our results indicate that the ensemble model is more robust, performing better than any individual base model trained on a specific feature extraction technique. A comparison of the mean, minimum, and standard deviation scores of three classification metrics (i.e., mean accuracy, precision and recall score on medium chain TEs) between EnZymClass and the five top performing base models is tabulated in Table 1. Detailed analysis of the 47 base models used in EnZymClass, each trained on a unique feature extraction technique is provided in **Table S3**. The results of a two-sided *t*-

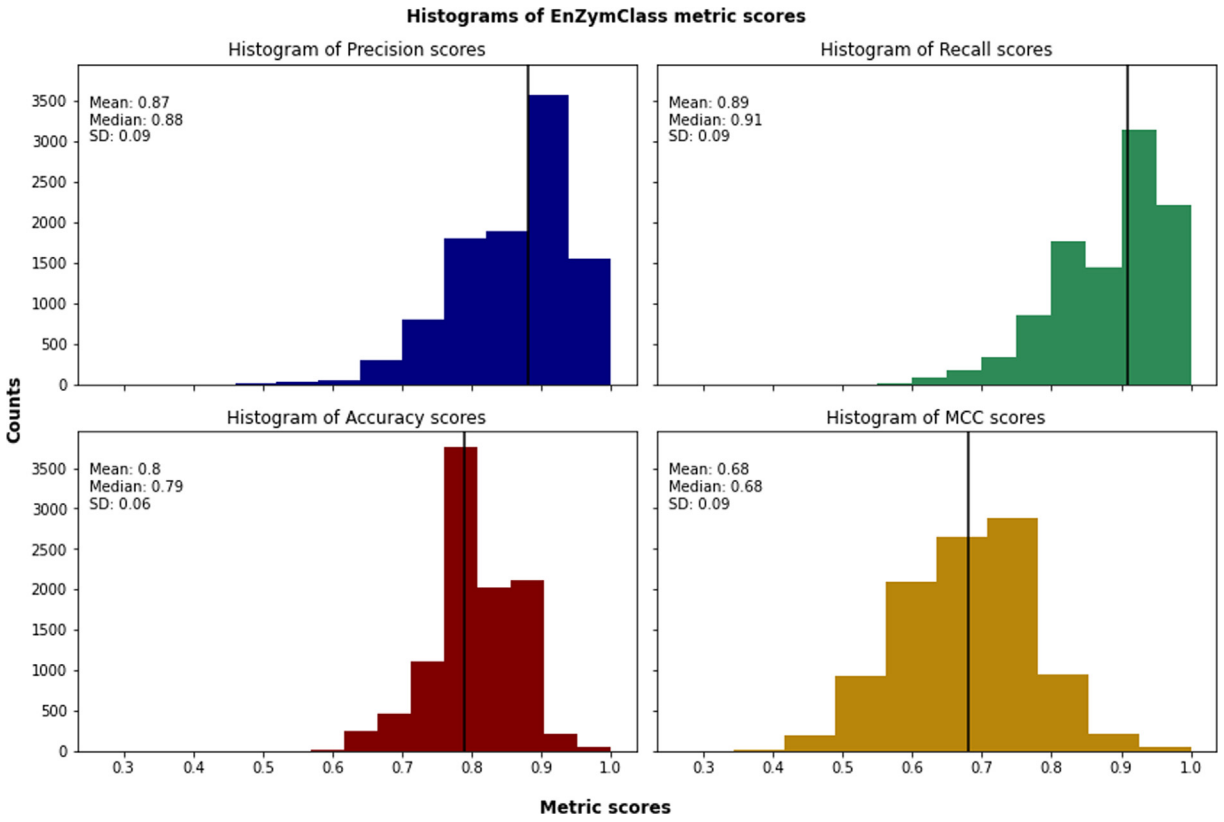

Fig. 2. Precision (on medium-chain TEs), Recall (on medium-chain TEs), Accuracy, and MCC score distribution of EnZymClass. The black vertical line denotes the median score of the specific metric. The mean, median, and worst precision scores are 0.87, 0.88, and 0.40, respectively. The mean, median, and worst recall scores are 0.89, 0.91, and 0.50, respectively. The mean, median, and worst accuracy scores are 0.8, 0.79 and 0.52, respectively. The mean, median, and worst MCC scores are 0.68, 0.68 and 0.27 respectively.

test that establishes the superiority of EnZymClass over any individual base learner in terms of accuracy and recall metrics is provided in **Table S6**.

3.3. Comparison with similarity-based classification method

We further benchmarked EnZymClass by comparing it to an existing sequence similarity-based classification method. Sequence similarity-based methods define a distance function to measure the similarity between a pair of sequences (Xing et al., 2010). Here, we calculated the BLASTP (Rédei, 2008) identity scores between a pair of subject and query sequence to use as the distance function. Henceforth, we trained a *k*-Nearest Neighbors classifier with *k* decided through cross-validation on a subset of TE sequences and predicted the substrate specificity on the remaining validation set. This process was repeated 10,000 times by varying the training and validation sets similar to the EnZymClass evaluation scheme. The distribution of the precision, recall, and accuracy scores for the similarity-based classification model along with the box plots comparing those distributions with similar validation score distributions achieved by EnZymClass is shown in Fig. 3. The results of a two-sided *t*-test that establishes the superiority of EnZymClass over the similarity model across all three metrics is provided in **Table S6**.

3.4. Identification of two uncharacterized medium-chain TEs from the Cuphea genus

EnZymClass predicted three enzyme sequences to encode medium-chain length specific TEs from the TE14 family in the ThYme database (Cantu et al., 2010). The TE14 family contained about 2,500 sequences of prokaryotic and eukaryotic sequences, however, since the training set consisted of solely eukaryotic sequences from the plant kingdom, the search was restricted to the subset of 617 sequences with eukaryotic origin. The *in vivo* performance of the predicted TEs when expressed in the *E. coli* RL08*ara* strain is summarized in Fig. 4**a** and **Table S8**. The three TEs, ClFatB3, ClFatB3-2, and CwFatB2 are from various *Cuphea* species: the former two from *Cuphea lanceolata* and the latter from *Cuphea wrightii*. ClFatB3 and CwFatB2 showed medium-chain length activity. ClFatB3-2 yielded a titer and distribution similar to the catalytically inactivated control, BTE-H204A. ClFatB3-2 differed from ClFatB3 by one amino acid in the binding pocket: position 135 in ClFatB3 is a serine while position 135 in ClFatB3-2 is a proline.

The free fatty acid distributions from the *in vivo* cultures contained a significant proportion of hexadecanoic and hexadecenoic acid. We reasoned that these species were cell membrane constituents which resulted from derivatization of phospholipids rather than free fatty acids. To validate this hypothesis, the cultures were centrifuged, and the supernatant was collected to be analyzed for free fatty acid content. The results from this processing step shown in Fig. 4**b** and **Table S9** confirmed that ClFatB3 and CwFatB2 are primarily medium-chain specific. While ClFatB3-2 did not exhibit medium-chain specificity, the analysis of the supernatant suggested this was due to enzyme inactivity. Since ClFatB3-2 only differed from ClFatB3 by a single residue, EnZymClass also classified ClFatB3 as a medium-chain specific TE. This instance illustrates the limitations of EnZymClass applied in this study, as it was tuned to only classify if a TE exhibits medium-chain, long-chain, or mixed substrate specificity. The addition of another classifier for predicting inactive TE could be in principle possible provided an inactive TE training dataset.

Table 1
EnZymClass performs better than any individual base model on varying validation datasets in terms of prediction accuracy and robustness to training set. We illustrate that phenomenon by comparing its results with that of five best performing base models (judging by their performance on validation datasets). The mean, minimum, and standard deviation of accuracy, precision, and recall scores achieved by EnZymClass and the five best performing base learners on varying validation datasets are displayed here.

Model Name	Mean Precision	Min Precision	Std Precision	Mean Recall	Min Recall	Std Recall	Mean Accuracy	Min Accuracy	Std Accuracy
EnZymClass	0.87	0.40	0.09	0.89	0.5	0.09	0.8	0.52	0.06
Spectrum Kernel	0.87	0.44	0.09	0.85	0.36	0.10	0.77	0.45	0.07
Gappy Kernel	0.87	0.44	0.09	0.86	0.38	0.10	0.77	0.48	0.07
CKSAAP	0.87	0.38	0.09	0.86	0.35	0.10	0.77	0.45	0.07
KSCTriad	0.86	0.33	0.09	0.85	0.33	0.10	0.76	0.41	0.07
Moran	0.87	0.33	0.10	0.85	0.38	0.10	0.76	0.45	0.08

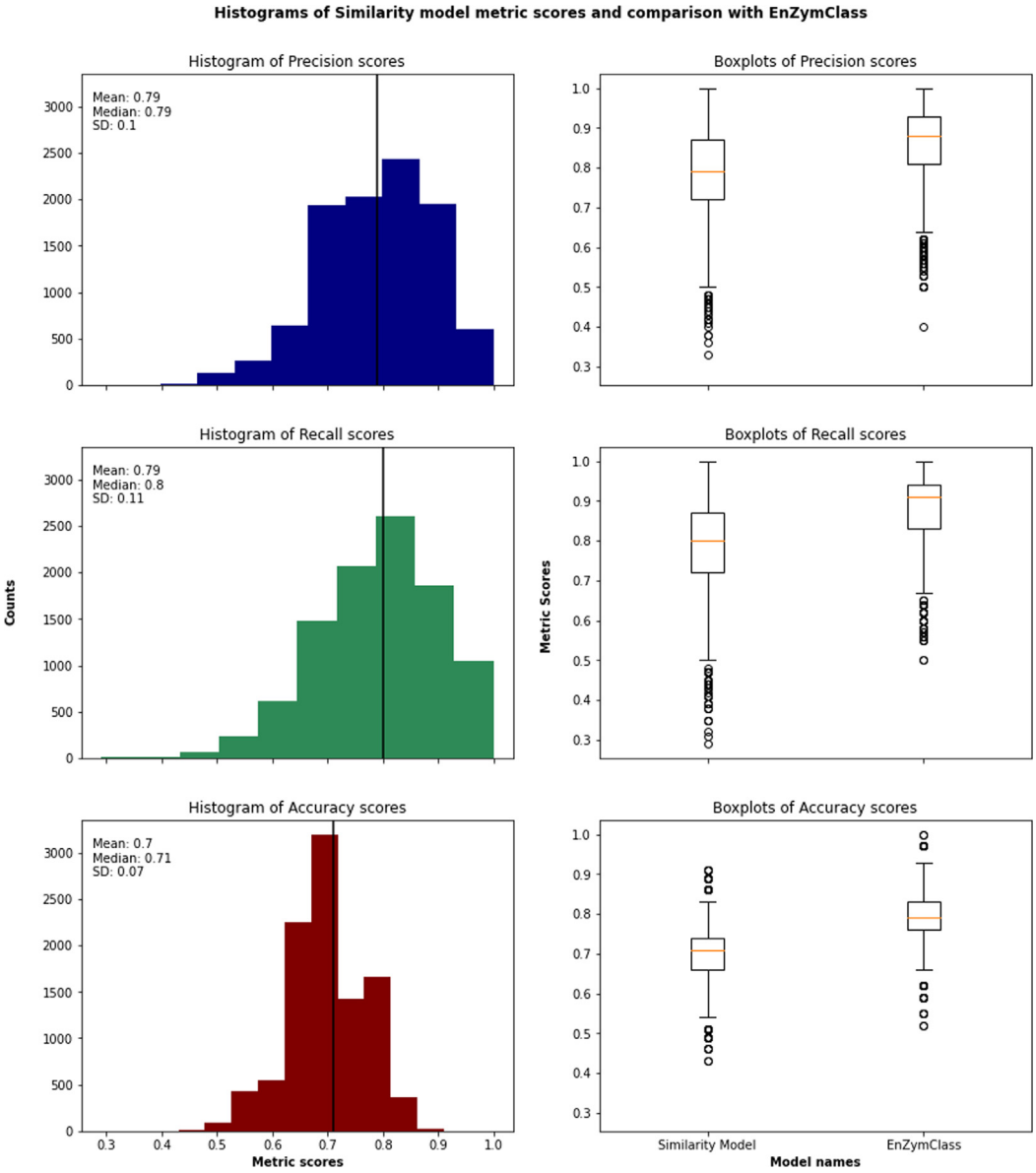

Fig. 3. Precision (on medium-chain TEs), Recall (on medium-chain TEs), and Accuracy score distribution of the similarity model along with the comparisons of those distributions with EnZymClass through box plots is depicted. The mean precision, recall, and accuracy scores are 0.79, 0.79, and 0.7, respectively. The box plot comparisons between Similarity model and EnZymClass performance metrics show the EnZymClass metric distributions are skewed towards higher scores across all metrics, thus asserting the superiority of EnZymClass over similarity-based model.

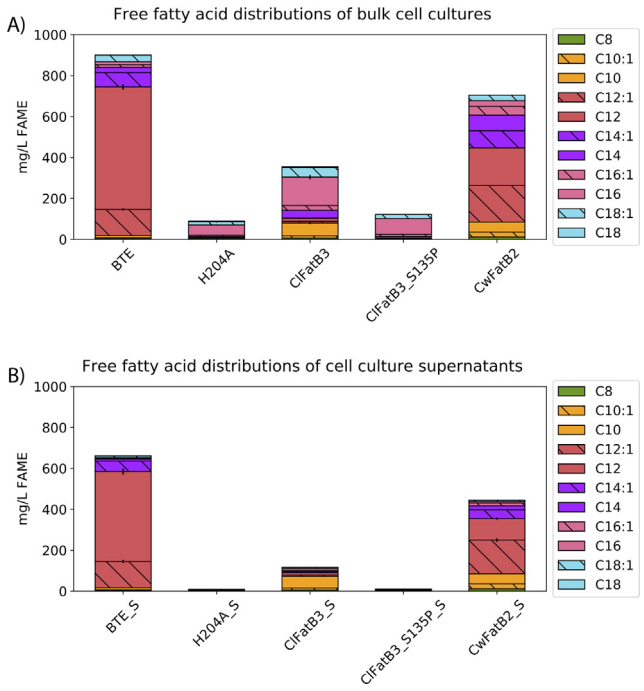

Fig. 4. The effect of TE homolog on free fatty acid distribution in a) bulk cultures and b) supernatants of RL08*ara E. coli* cells. The distribution from cells expressing the California bay laurel TE (BTE) and a catalytically inactive BTE variant (H204A) are shown as a positive and negative control, respectively. Bar height represents the average titer obtained from biological triplicates, and error bars represent the standard error of the mean.

3.5. Improving activity of medium-chain acyl-ACP TE identified from ML predictions

We sought to improve the activity of ClFatB3 and CwFatB2 enzymes by incorporating mutations previously demonstrated to enhance free fatty acid production. Hernández-Lozada et al. leveraged an auxotrophic *E. coli* strain to screen for mutations productive toward enhancing activity of an octanoyl-ACP TE from *Cuphea palustris*, CupTE (Hernández Lozada et al., 2018). The three mutations which conferred the increased activity included a truncation, a mutation near the N-terminus, and a mutation in the binding pocket (ΔA_{54}, N28S, and I65M, respectively). The truncation was first incorporated into all three TE sequences identified from the model predictions as well as to the BTE sequence. The effect of this truncation on free fatty acid distribution is shown in **Fig. S4** and in **Table S8**. The Clustal Omega multiple sequence alignment used to locate the truncation site is shown in **Fig. S5** (Larkin et al., 2007).

The N-terminal truncation resulted in a 3.3-fold improvement in C_{10} free fatty acid titer. The amino acid substitutions identified in the active CupTE variant were incorporated into the truncated ClFatB3 sequence by cloning the D10S and I47M residue changes into the expression vector (see **Fig. S5**). The effect of these substitutions as single point mutations and in combination is shown in **Fig. S6**. The D10S point mutation was shown to have the best effect in improving the overall titer in production cultures; this modification resulted in an 18% improvement in decanoic and decenoic acid titer compared to the truncated ClFatB3 variant and a 4.2-fold improvement when compared to the WT. This final ClFatB3 variant also resulted in a product distribution of 70% of the ten-carbon species, excluding membrane lipid chain-lengths. The efforts in this study could be coupled with subsequent mutagenesis efforts to tailor substrate specificity, ultimately yielding a highly active and highly specific decanoyl-ACP TE.

EnZymClass was used to successfully identify two TE in the TE14 family in the ThYme database which exhibit medium-chain-length activity in *E. coli*. These enzymes, ClFatB3 and CwFatB2, both originated from the *Cuphea* genus and resulted in an *in vivo* product profile of decanoic acid and dodecanoic acid, respectively. Since both of these discoveries originated from a plant genus known to have medium-length acyl chains in its seed triglycerides, we wanted to ensure that EnZymClass was not simply returning sequences from *Cuphea* hosts and indeed was able to discern between medium and long-chain acyl-ACP TE homologs. To explore this, we tested ClFatB4, another TE from *Cuphea lanceolata* which was in the TE14 ThYme database family. We also tested CpaFatB1, CpaFatB2A, and CpaFatB3, TEs from *Cuphea paucipetala*, to sample other enzymes from the *Cuphea* genus. The free fatty acid distribution from these homologs were all mostly dominated by the longer chemical species, as predicted by EnZymClass. This confirmed EnZymClass' capability to classify substrate specificity among homologs from the same species **(Fig. S7 and Table S8).**

4. Discussion

We presented a computational tool, termed EnZymClass, designed to address barriers to accurate classification of protein subfunction from primary sequence information. To test the effectiveness of EnZymClass, we applied it to a specific protein classification task: categorization of plant acyl-ACP TEs by substrate specificity. When provided full-length amino acid sequences of putative acyl-ACP TE from the plant kingdom, EnZymClass can flag enzymes with the potential to produce distributions of over 50% of medium-chain free fatty acids in an *E. coli* production strain. We used EnZymClass to characterize all eukaryotic acyl-ACP TEs in the ThYme database (Cantu et al., 2010). Among the three TEs predicted to be medium-chain specific by EnZymClass, two were experimentally validated to possess the desired activity.

To ratchet the activity of the TE discoveries made from EnZymClass' predictions, the ClFatB3 and CwFatB2 genes were modified according to previous successful TE engineering efforts. Most notably, the final ClFatB3 variant, ClFatB3_trunc_M1, resulted in a 4.2-fold overall improvement in C_{10} free fatty acid titer compared to the WT enzyme. We obtained this enhancement by incorporating the corresponding mutations described by Hernández-Lozada et al. which led to activity improvement in the acyl-ACP TE from *Cuphea palustris*. The rationale for this observed performance boost in the TEs from *Cuphea palustris* and *Cuphea lanceolata* (ClFatB3) but not in *Cuphea wrightii* (CwFatB2) remains unknown; these mutations do not occur inside the binding pocket or on the ACP landing pad, the two most studied mechanisms for changes in TE activity and specificity (Jing et al., 2018; Sarria et al., 2018). Nonetheless, since the truncation did not benefit CwFatB2 the same extent exhibited by ClFatB3, we reasoned incorporating the other amino acid substitutions from Hernández-Lozada et al. would not be as fruitful as well. Future efforts will entail the mutagenesis of ClFatB3 for enhanced substrate specificity to achieve near exclusive selectivity for C_{10} FFA production in bacterial hosts. With this work, we demonstrate how ML can be implemented to facilitate an enzyme engineering pipeline.

Previous studies indicated that sequence similarity is not the best predictor of TE substrate specificity since highly similar sequences may have different substrate specificity (Jing et al., 2011, 2018). For example, the TEs from *Cuphea viscosissima* (NCBI accession numbers AEM72522.1 and AEM72523.1) share more than 70% sequence identity but display different substrate specificities. In addition, TEs from *Ulmus americana* and *Umbellularia californica* (NCBI accession numbers AAB71731 and AAA34215.1) both display medium-chain acyl-ACP substrate specificity yet share less than 52% sequence similarity. This underscores the utility of feature extraction techniques from protein sequences used in EnZymClass, which not only encode similarity

between homologous sequences but also extract physicochemical, contextual, and evolutionary information from protein sequences.

Although we have attained reasonably high accuracy on TE substrate specificity classification task, we acknowledge that EnZymClass is currently unable to accomplish a deeper level of TE classification across specific chain lengths (*C8, C10, C12, C14, C16, C16:1, C18, and C18:1* or 8-class classification). It must be noted that the referred limitation to achieve a higher resolution of TE classification can be largely ascribed to a lack of characterized TE dataset with enough instances for each chain length. Furthermore, we recognize that the use of PCA to reduce feature dimensionality inhibits extraction of insightful features that contribute towards substrate specificity. Although PCA helps us to retain model generalizability, we are currently working on substituting the PCA step with an alternate dimensionality reduction technique which will allow insightful feature extraction. Additionally, we plan to incorporate embedding based feature extraction techniques, such as UniRep (Alley et al., 2019), in future improvements of EnZymClass. However, fine-tuning is essential for embedding based feature extraction techniques to achieve superior performance. The dataset used in this work does not have enough instances to allow fine-tuning for deep learning models. We anticipate that in the future as larger datasets become available, we will be able to make use of such methodological enhancements.

EnZymClass can be adapted to other protein classification challenges ranging from general tasks such as protein structural class prediction or protein-protein interactions to more defined such as TE substrate specificity prediction or protein glycosylation site prediction (Chauhan et al., 2012). While general applications such as protein-protein interactions suffer from dataset imbalance (Yu et al., 2010), more specific tasks, for instance glycosylation site prediction may encounter yet another set of difficulties relating to small sized datasets. Issues related to high dimensionality and correlated feature set are ubiquitous in the protein classification domain. EnZymClass partially alleviates protein classification challenges while maintaining the computational efficiency required for swift functional characterization.

CRediT authorship contribution statement

Deepro Banerjee: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Michael A. Jindra:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Alec J. Linot:** Conceptualization, Methodology. **Brian F. Pfleger:** Resources, Writing – original draft, Supervision, Project administration, Funding acquisition. **Costas D. Maranas:** Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Science Foundation (CBET-1703504, link: https://www.nsf.gov/) and the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE- SC0018420, link: https://cabbi.bio/), both received by BFP and CDM. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy or the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The sequences for *Auxenochlorella protothecoides* (KFM28838.1) and *Prunus sibirica* L. (AIX97815.1) were synthesized by the U.S. Department of Energy Joint Genome Institute (JGI), a DOE Office of Science User Facility, supported by the Office of Science of the U.S. Department of Energy. Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crbiot.2021.12.002.

References

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M., 2019. Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods 16 (12), 1315–1322. https://doi.org/10.1038/s41592-019-0598-1.

Amin, S.R., Erdin, S., Ward, R.M., Lua, R.C., Lichtarge, O., 2013. Prediction and experimental validation of enzyme substrate specificity in protein structures. Proc. Natl. Acad. Sci. 110 (45), E4195–E4202. https://doi.org/10.1073/pnas.1305162110.

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B., Rätsch, G., 2008. Support vector machines and kernels for computational biology. PLoS Comput. Biol.. https://doi.org/10.1371/journal.pcbi.1000173.

Çamoğlu, Orhan, Can, Tolga, Singh, A.K., Wang, Yuan-Fang, 2005. Decision tree based information integration for automated protein classification. J. Bioinformat. Comput. Biol. 03 (03), 717–742. https://doi.org/10.1142/S0219720005001259.

Cantu, D.C., Chen, Y., Reilly, P.J., 2010. Thioesterases: A new perspective based on their primary and tertiary structures. Protein Science, vol. 19, issue 7, pp. 1281–1295. https://doi.org/10.1002/pro.417.

Chauhan, J.S., Bhat, A.H., Raghava, G.P.S., Rao, A., 2012. GlycoPP: A webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. PLoS ONE 7 (7). https://doi.org/10.1371/journal.pone.0040155.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.-C., Song, J., Valencia, A., 2018. IFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34 (14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140.

Dehesh, K., Jones, A., Knutzon, D.S., Voelker, T.A., 1996. Production of high levels of 8:0 and 10:0 fatty acids in transgenic canola by overexpression of Ch FatB2, a thioesterase cDNA from Cuphea hookeriana. Plant J. 9 (2), 167–172. https://doi.org/10.1046/j.1365-313X.1996.09020167.x.

Deshpande, M., Karypis, G., 2002. Evaluation of techniques for classifying biological sequences. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/3-540-47887-6_41.

Ding, C.H.Q., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17 (4), 349–358. https://doi.org/10.1093/bioinformatics/17.4.349.

Gordon Roessler, P., Roy, G., 2015. ACYL-ACP THOESTERASE GENES AND USES THEREFOR (Patent No. 8956834 B2).

Greenhalgh, J., Saraogee, A., Romero, P.A., 2018. Data-driven protein engineering. In: Unknown.

Grisewood, M.J., Hernández-Lozada, N.J., Thoden, J.B., Gifford, N.P., Mendez-Perez, D., Schoenberger, H.A., Allan, M.F., Floy, M.E., Lai, R.-Y., Holden, H.M., Pfleger, B.F., Maranas, C.D., 2017. Computational Redesign of Acyl-ACP Thioesterase with Improved Selectivity toward Medium-Chain-Length Fatty Acids. ACS Catal. 7 (6), 3837–3849.

Hernández Lozada, N.J., Lai, R.-Y., Simmons, T.R., Thomas, K.A., Chowdhury, R., Maranas, C.D., Pfleger, B.F., 2018. Highly Active C 8 -Acyl-ACP Thioesterase Variant Isolated by a Synthetic Selection Strategy. ACS Synth. Biol. 7 (9), 2205–2215. https://doi.org/10.1021/acssynbio.8b0021510.1021/acssynbio.8b00215.s001.

Hernández Lozada, N.J., Simmons, T.R., Xu, K., Jindra, M.A., Pfleger, B.F., 2020. Production of 1-octanol in Escherichia coli by a high flux thioesterase route. Metab. Eng. 61 (April), 352–359. https://doi.org/10.1016/j.ymben.2020.07.004.

Jing, F., Cantu, D.C., Tvaruzkova, J., Chipman, J.P., Nikolau, B.J., Yandeau-Nelson, M. D., Reilly, P.J., 2011. Phylogenetic and experimental characterization of an acyl-ACP thioesterase family reveals significant diversity in enzymatic specificity and activity. BMC Biochem. 12 (1). https://doi.org/10.1186/1471-2091-12-44.

Jing, F., Zhao, L., Yandeau-Nelson, M.D., Nikolau, B.J., 2018. Two distinct domains contribute to the substrate acyl chain length selectivity of plant acyl-ACP thioesterase. Nat. Commun. 9 (1), 860. https://doi.org/10.1038/s41467-018-03310-z.

Jones, A., Davies, H.M., Voelker, T.A., 1995. Palmitoyl-Acyl Carrier Protein (ACP) Thioesterase and the Evolutionary Origin of Plant ACyl-ACP Thioesterases. Plant Cell 7 (March), 359–371.

Jurafsky, D., Martin, J.H., 2016. Language Modeling with N- grams. Speech Langu. Process..

Khurana, P., Gokhale, R.S., Mohanty, D., 2010. Genome scale prediction of substrate specificity for acyl adenylate superfamily of enzymes based on active site residue profiles. BMC Bioinf. 11 (1). https://doi.org/10.1186/1471-2105-11-57.

Kim, H.J., Silva, J.E., Vu, H.S., Mockaitis, K., Nam, J.-W., Cahoon, E.B., 2015. Toward production of jet fuel functionality in oilseeds: Identification of FatB acyl-acyl carrier protein thioesterases and evaluation of combinatorial expression strategies in Camelina seeds. J. Exp. Bot. 66 (14), 4251–4265. https://doi.org/10.1093/jxb/erv225.

Kim, S., Clomburg, J.M., Gonzalez, R., 2015. Synthesis of medium-chain length (C6–C10) fuels and chemicals via β-oxidation reversal in Escherichia coli. J. Ind. Microbiol. Biotechnol. 42 (3), 465–475. https://doi.org/10.1007/s10295-015-1589-6.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23 (21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404.

Lennen, R.M., Braden, D.J., West, R.M., Dumesic, J.A., Pfleger, B.F., 2010. A process for microbial hydrocarbon synthesis: Overproduction of fatty acids in Escherichia coli and catalytic conversion to alkanes. Biotechnol. Bioeng. 106 (2), 193–202. https://doi.org/10.1002/bit.22660.

Lennen, R.M., Pfleger, B.F., 2012. Engineering Escherichia coli to synthesize free fatty acids. Trends Biotechnol. 30 (12), 659–667. https://doi.org/10.1016/j.tibtech.2012.09.006.

Leslie, C., Eskin, E., Noble, W.S., 2002. The spectrum kernel: a string kernel for SVM protein classification. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. https://doi.org/10.1142/9789812799623_0053.

Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S., 2004. Mismatch string kernels for discriminative protein classification. Bioinformatics 20 (4), 467–476. https://doi.org/10.1093/bioinformatics/btg431.

Nanni, L., Lumini, A., Brahnam, S., 2013. An empirical study on the matrix-based protein representations and their combination with sequence-based approaches. Amino Acids 44 (3), 887–901. https://doi.org/10.1007/s00726-012-1416-6.

Nanni, L., Lumini, A., Brahnam, S., 2014. An empirical study of different approaches for protein classification. Sci. World J. 2014, 1–17. https://doi.org/10.1155/2014/236717.

Noweck, K., Ridder, H., 1988. Fatty Alcohols - Industrial Production. In: Ullmann's encyclopedia of industrial chemistry. 5th ed. Wiley-VCH, pp. 277–295.

Ohlrogge, J., Thrower, N., Mhaske, V., Stymne, S., Baxter, M., Yang, W., Liu, J., Shaw, K., Shorrosh, B., Zhang, M., Wilkerson, C., Matthäus, B., 2018. PlantFAdb: a resource for exploring hundreds of plant fatty acid structures synthesized by thousands of plants and their phylogenetic relationships. Plant J. 96 (6), 1299–1308. https://doi.org/10.1111/tpj.14102.

Palme, J., Hochreiter, S., Bodenhofer, U., 2015. KeBABS: An R package for kernel-based analysis of biological sequences. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv176.

Pfleger, B. F., Hernandez-Lozada, N., Lai, R.-Y., 2020. Mutant thioesterases (Patent No. US10844410B2).

Pfleger, B.F., Lennen, R.M., 2013. Fatty acid-producing hosts (Patent No. US8617856B2).

Politz, M., Lennen, R., Pfleger, B., Engineering, B., 2016. Quantification of Bacterial Fatty Acids by Extraction and Methylation. Bio Protocols 3 (21).

PyPI, 2020. PyPI · The Python Package Index. PyPI.

Rédei, G.P., 2008. BLASTP. Encycl. Genet. Genom. Proteom. Informat.. https://doi.org/10.1007/978-1-4020-6754-9_1881.

Rigouin, C., Croux, C., Borsenberger, V., Ben Khaled, M., Chardot, T., Marty, A., Bordes, F., 2018. Increasing medium chain fatty acids production in Yarrowia lipolytica by metabolic engineering. Microb. Cell Fact. 17 (1). https://doi.org/10.1186/s12934-018-0989-5.

Rupilius, W., Ahmad, S., 2007. Palm oil and palm kernel oil as raw materials for basic oleochemicals and biodiesel. Eur. J. Lipid Sci. Technol. 109 (4), 433–439. https://doi.org/10.1002/ejlt.200600291.

Saigo, H., Vert, J.-P., Ueda, N., Akutsu, T., 2004. Protein homology detection using string alignment kernels. Bioinformatics 20 (11), 1682–1689. https://doi.org/10.1093/bioinformatics/bth141.

Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Bioinf. 9 (1), 56–68. https://doi.org/10.1002/prot.340090107.

Sarria, S., Bartholow, T.G., Verga, A., Burkart, M.D., Peralta-Yahya, P., 2018. Matching Protein Interfaces for Improved Medium-Chain Fatty Acid Production [Rapid-communication]. ACS Synth. Biol. 7 (5), 1179–1187. https://doi.org/10.1021/acssynbio.7b0033410.1021/acssynbio.7b00334.s001.

Sarria, S., Kruyer, N.S., Peralta-Yahya, P., 2017. Microbial synthesis of medium-chain chemicals from renewables. Nat. Biotechnol. 35 (12), 1158–1166. https://doi.org/10.1038/nbt.4022.

Wang, J., Yang, B., Revote, J., Leier, A., Marquez-Lago, T.T., Webb, G., Song, J., Chou, K.-C., Lithgow, T., Hancock, J., 2017. POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics 33 (17), 756–2758. https://doi.org/10.1093/bioinformatics/btx302.

Whalen, S., Pandey, G., 2013. A comparative analysis of ensemble classifiers: Case studies in genomics. In: Proceedings - IEEE International Conference on Data Mining, ICDM. https://doi.org/10.1109/ICDM.2013.21.

Xing, Z., Pei, J., Keogh, E., 2010. A brief survey on sequence classification. ACM SIGKDD Explorat. Newsletter 12 (1), 40–48.

Xu, P., Qiao, K., Ahn, W.S., Stephanopoulos, G., 2016. Engineering Yarrowia lipolytica as a platform for synthesis of drop-in transportation fuels and oleochemicals. PNAS 113 (39), 10848–10853. https://doi.org/10.1073/pnas.1607295113.

Yang, K.K., Wu, Z., Arnold, F.H., 2019. Machine-learning-guided directed evolution for protein engineering. Nat. Methods 16 (8), 687–694. https://doi.org/10.1038/s41592-019-0496-6.

Yu, C.-Y., Chou, L.-C., Chang, D.-H., 2010. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. BMC Bioinf. 11 (1). https://doi.org/10.1186/1471-2105-11-167.

Yuan, L., Voelker, T.A., Hawkins, D.J., 1995. Modification of the substrate specificity of an acyl-acyl carrier protein thioesterase by protein engineering. Proc. Natl. Acad. Sci. 92 (23), 10639–10643. https://doi.org/10.1073/pnas.92.23.10639.