## Introduction

\*\*Introduction about Thioesterases and their importance\*\*.

A key value proposition of synthetic biology is providing access to chemicals which are not sustainably produced at commercial scales. Medium-chain oleochemicals, 8 to 12-carbon free fatty acids and derivatives, are one such class of products. While these chain lengths have traditionally been sourced from the tropical crops, such as palm, palm kernel, and coconut, the 8, 10, and 12-carbon products are not major constituents of the oil [1]. Furthermore, the displacement of rainforest habitat due to the cultivation of the oil palm has been identified as having the single largest impact on decreasing biodiversity observed in the Southeast Asian jungle ecosystem [2]. Processes have been established to create the 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 [3].

\*\*Connection between your Introduction with mine: experimental methods are expensive; ML or other computational methods can be cost-effective. \*\*

As an alternative, the field of synthetic biology has achieved fatty acid and fatty alcohol distributions with over 90% of the product belonging to the C8 species [4], [5]. This has been achieved via rewiring of the fatty acid biosynthesis pathway in *E. coli*, namely by the incorporation of an engineered 8-carbon specific acyl-ACP Thioesterase (TE) from *Cuphea palustris*. 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* production systems [4], [5], [6], [7], [8], [9] **(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 [7], [10], [11], [12]. Thus, several efforts have been made to bioprospect genomes of plants with high fractions of the medium-chain oils to identify and implement the TE gene responsible for the narrow substrate specificity [13], [4]. While progress has been demonstrated in identifying the features which dictate specificity in acyl-ACP TEs among plants [14], the throughput for bioprospecting, characterizing, and in some cases, engineering the acyl-ACP TE is largely inhibited by the testing pipeline, which requires derivatization of the free fatty acids into fatty acid methyl esters prior to analysis with gas chromatography [15]. A method for inferring substrate specificity from TE gene sequence would therefore expedite this process, removing the necessity of expressing each homolog in a host to gain insight to its selectivity profile. Jing *et al.* showed that phylogenetic and sequence identity analysis alone were not sufficient to distinguish plant TE substrate specificity [8].

Machine learning (ML) has been demonstrated to effectively classify proteins into different functional classes based on their primary sequence information. While some of these approaches have also incorporated structural information, ML algorithms have successfully identified pertinent information from primary sequence to distinguish between highly similar proteins: guanylyl and adenylyl cylases, lactate and malate dehydrogenases, trypsins and chymotrypsins [16]. A variety of approaches to solve the protein classification problem have been proposed, which broadly fall under two categories, generative and discriminative. The generative approach builds a model of the feature distribution for each protein category and assigns a particular class to a candidate protein sequence by evaluating how well the sequence fits the model. Methods that fall under the generative category are based on sequence similarity comparison using local alignment similarity scores [17], [18], position specific scoring tables or profiles created from a group of previously aligned sequences [19], [20], consensus sequence patterns or motifs [21] and Hidden Markov Models (HMMs) [22]–[24]. However, generative approaches are either highly dependent on the database used to search for sequence similarity (local alignment and profile based similarity search) [25] or are computationally expensive (HMMs) [26]. The discriminative approach on the other hand focuses on accurately learning the decision boundary between classes. Commonly used discriminative approaches rely on training classifiers like Support Vector Machine (SVMs) or Neural Network (NNs) to learn discriminative rules from both positive (belonging to a particular protein class) and negative (not belonging to that protein class) set of protein sequences and using the learnt rules to predict the class of any new protein sequence [27]–[29].

Recent results suggest that discriminative approaches have outperformed generative approaches both in terms of accuracy and computational efficiency of solving the protein classification problem [30]. SVM is among the most widely used discriminative learning algorithm which has been proven to be extremely effective on sequence based classification [27]–[34]. SVM has been used to achieve state of the art performance to detect remote protein homologies and classify proteins in the SCOP database into major structural classes [26], [27]. The performance of an SVM classifier is highly influenced by the feature extraction technique employed to encode the protein sequences [35].

Feature extraction of protein sequences aims at formulating a discrete numerical representation of a protein to create feature vectors that are correlated with the desired attribute of the protein one would like to predict. In order to train an SVM, a number of feature extraction techniques for protein sequences have been suggested in literature which can be divided into two categories, kernel based methods and vector representation of protein sequences [36]. The Fisher kernel introduced by Jakkola et. al. is one of the first kernel based feature extraction technique used to classify proteins based on their sequence information [37]. It was followed by the spectrum kernel [28] and its more generalized form, the mismatch kernel [30], both introduced by Leslie et. al. which achieved similar performances in terms of accuracy when compared to the Fisher kernel but is computationally much less expensive. The weighted degree kernel introduced by Ratsch et. al. also took the position of the substrings within the protein sequence into account as opposed to the spectrum and mismatch kernel introduced by Leslie et. al. [38]. It was used to identify alternatively spliced exons in C. elegans. Apart from kernel based methods, the second class of feature representation technique extracts structural and physicochemical properties embedded in the protein sequence and converts it into a numerical vector. One of the first and simplest discrete model to represent protein sequences that falls under the second class is Amino Acid Composition (AAC) developed by Nakashima et. al. which was used to classify proteins into different folding types with high accuracy [39]. Dubchak et. al. developed the more complicated Composition-Transition-Distribution (CTD) descriptor that takes into account different physical and stereochemical properties of the amino acids in the protein sequences like amino acid composition, predicted secondary structure, hydrophobicity, normalized van der waals volume, polarity and polarizability to construct the feature vector [27]. Chou et. al. upgraded the simple AAC encoder developed by Nakashima et. al. to a pseudo Amino Acid (pseAA) composition encoder that is able to retain some pattern specific information embedded in the protein sequence [40]. Features have also been derived from Position Specific Scoring Matrices (PSSMs) profiles, which contains evolutionary information [41]. While selection of the most informative feature extraction technique has resulted in an improved performance of a classifier, the use of ensemble methods which combines the output of multiple classifiers has also helped to attain greater accuracy while solving the classification problem [42].

Several studies have shown that ensemble methods performed better than any individual classification method especially in problems relevant to the protein classification domain [43]–[45]. Camoglu et. al used a decision tree based ensemble classifier to classify protein in the SCOP database and showed how it is possible to attain much lower error rates using the ensemble classifier than any individual method [43]. Diplaris et. al. performed an empirical study where they compared the performance of several individual algorithms to solve the motif based classification problem and demonstrated the positive effect of combining different classification algorithms on prediction accuracy [44]. Tan et. al. illustrated the advantage of using ensemble classifiers on imbalanced datasets while solving the protein fold classification problem [45]. Similarly, Caragea et. al. trained an ensemble of SVM classifiers to predict glycosylation sites in amino acid residues and found that an ensemble of SVMs outperformed an individual SVM trained on imbalanced and balanced data [46].

The goal of this study is to classify plant acyl-ACP Thioesterases (TE) based on their substrate specificity. Information about characterized TEs were collected as a part of this study from multiple literature sources [7], [47]–[49] and they were categorized into three different classes in order to formulate a multi-class classification problem. To solve the classification problem, a stacked ensemble framework was developed that comprises of three base learners trained using three different feature extraction techniques from protein primary sequences and a meta learner which combines the output of the three base learners by applying a majority voting criterion to predict the substrate specificity class of TEs. The purpose of using an ensemble method was to deal with the imbalanced dataset and get high prediction accuracy. Our results illustrate the advantage of using an ensemble method with imbalanced dataset compared to any individual method, in accordance with previous studies [45], [46]. The base learners in the ensemble are all SVM classifiers similar to the works of Caragea et. al. [46] and Nanni et. al. [36] where they have used an ensemble of SVMs to boost prediction accuracy. SVM was selected as the primary learning algorithm because of their ability to deal with high dimensional dataset and generalizability [50]. We have also experimented with more complex models like Neural Network and Gradient Boosted Trees as the base learner, but they were both outperformed by SVM in terms of accuracy on separate held-out validation set. The proteins were represented by their amino acid sequences. The three feature extraction techniques used were the spectrum kernel introduced by Leslie et. al. [28], a variant of the spectrum kernel, where the amino acids were first grouped into 6 categories to create a compressed representation based on their physicochemical attributes and the spectrum kernel was used to extract features from this compressed representation, and a newly introduced feature representation technique, autopos detection, which automatically detects the positions within the protein sequence which are maximally correlated with the functional attribute of the proteins. These three different feature representation techniques were used to train the three base models in the ensemble. We have also analyzed the effect of using 21 other representation techniques which extracts physicochemical attributes from protein sequences like CTD,… [27] but they lowered the performance of the ensemble. The feature vectors generated through the extraction process were decomposed into lower dimensional and linearly uncorrelated features using Principal Component Analysis. The reduction in dimensionality of the feature vectors was performed to prevent overfitting. The lower dimensional and decomposed set of feature vectors were used to train the classifiers and predict substrate specificity of TEs. The ensemble method achieved a mean validation accuracy of 0.76 across 10,000 simulations of this study using different training and validation sets. However, the worst case accuracy across simulations was 0.45 which indicates that the method is not extremely robust to the training set. One possible reason behind the lack of robustness could be that the decision boundaries between the protein classes is not well defined. SVMs are maximal margin classifiers and the thickness of the decision boundary plays a very important role governing the robustness of the model [51]. The mean precision score of the model across the simulations for the medium chained TEs, the product of interest, was 0.84. Hence, we shifted to a regression based framework, where instead of classifying the proteins into three separate categories and assigning discrete labels based on their substrate specificity, continuous values labels were attributed to each protein. \*\*The labels were generated\*\*. The feature vectors for the protein sequences in the regression framework were generated through the grouped amino acid spectrum kernel method followed by PCA. An individual SVM was used as the regression model which achieved an average mean squared error of 0.0348 on held-out cross validation set across 10,000 simulations. The worst performing model achieved a mean squared error of 0.1067.

\*\*bioprospecting part\*\*.

We hypothesize that bioprospecting for novel, uncharacterized, medium-chain thioesterases could be facilitated by using machine learning to predict substrate specificity from gene sequence. To test this hypothesis, we trained a SVM with 116 thioesterase sequences previously characterized in *E. coli*. We intend to use this model to identify a 10-carbon specific acyl-ACP TE among a set of uncharacterized TE enzymes from select plants known to have predominantly decanoyl chains in their seed oils. This study may show that small to medium-sized datasets can be leveraged to guide bioprospecting efforts while simultaneously supporting the endeavor of the synthetic biology community to provide access to chemicals not easily obtained through conventional methods.