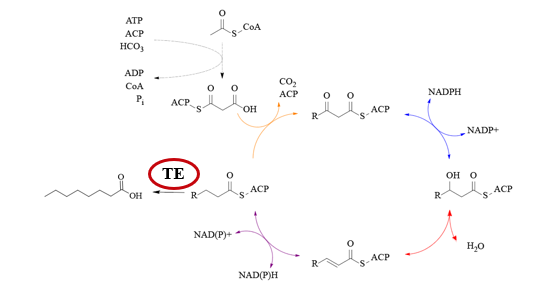
# Introduction

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-carbons 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 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].

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 from *Cuphea palustris*. Indeed, the expression of various acyl-ACP thioesterases, 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] **(Fig. 1)**. Of these studies, acyl-ACP thioesterases from select plant species have been shown have greater native specificity toward the medium-chain substrates when compared to bacterial homologs [7], [9]. Thus, several efforts have been made to bioprospect genomes of plants with high fractions of the medium-chain oils to identify and implement the thioesterase gene responsible for the narrow substrate specificity [10], . While progress has been demonstrated in identifying the features which dictate specificity in acyl-ACP thioesterases among plants [11], the throughput for bioprospecting, characterizing, and in some cases, engineering the acyl-ACP thioesterase 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 [12]. A method for inferring substrate specificity from thioesterase gene sequence would therefore expedite this process, removing the necessity of expressing each homolog in a host to gain insight to its selectivity profile. We hypothesized that bioprospecting for novel, uncharacterized, medium-chain thioesterases could be facilitated by using machine learning to predict substrate specificity from gene sequence. Our goal was to apply the machine learning algorithm to identify a 10-carbon specific acyl-ACP thioesterase among a set of uncharacterized thioesterases from select plants known to have predominantly decanoyl chains in their seed oils, thus testing the aforementioned hypothesis while simultaneously supporting the endeavor of the synthetic biology community to provide access to chemicals not easily obtained through conventional methods.



**Figure 1:** The acyl-ACP thioesterase plays a key role in fatty acid biosynthesis in *E. coli*. By intercepting the growing acyl-ACP chains, the thioesterase hydolyzes 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*.

# Methods

We have used three different feature representation techniques in our formulation which resulted in a diverse set of individual models required by the ensemble to capture various trends in our dataset. Each of the techniques is described as follows:

## k-mer motif builder

## GAAC based k-mer motif builder

## Binding pocket specific positional model

Our base model creates a PCA based encoding of the feature space and then uses SVM to predict the enzyme classes.