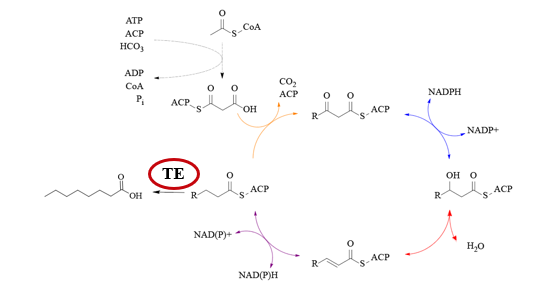
# Introduction

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-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 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], [9] **(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], [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 thioesterase gene responsible for the narrow substrate specificity [13], [4]. While progress has been demonstrated in identifying the features which dictate specificity in acyl-ACP thioesterases among plants [14], 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 [15]. 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. Jing *et al.* showed that phylogenetic and sequence identity analysis alone were not sufficient to distinguish plant thioesterases 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 sequence information to distinguish between highly similar proteins: guanylyl and adenylyl cylases, lactate and malate dehydrogenases, trypsins and chymotrypsins [16].

We hypothesized 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 then used this model to identify a 10-carbon specific acyl-ACP thioesterase among a set of uncharacterized thioesterase enzymes from select plants known to have predominantly decanoyl chains in their seed oils. This study shows 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.



**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

### Dataset Compilation

The training dataset included primary sequence and accompanying *in vivo* *E. coli* product distributions for 116 acyl-ACP plant TEs previously reported in scientific and patent literature. [Direct to supplementary table which is currently Thioesterase\_Master\_List.xlsx]. *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 thioesterase using discrete categories and a regression framework. For models which used discrete classification, the TE were grouped into one of three categories. The “medium-chain” category contained TE which resulted in distributions of at least 50% C8 to C12 free fatty acids. The “long-chain” category contained TE which produced 50% C14 to C18 free fatty acids and less than 10% C8 to C12 free fatty acids. Finally, the “mixed distribution” category contained TE which yielded distributions between 10% and 50% C8 to C12 free fatty acids. In the implementation of the linear regression classifier, each sequence was assigned a number which represented the fraction of the total free fatty acid distribution constituted of C8 to C12 free fatty acids.

Feature Selection for ML Prediction of ACP Thioesterase Specificity

A total of 43 residue positions were selected for feature encoding for ML prediction of acyl-ACP thioesterase specificity. These 43 residues were selected based on their importance in prior acyl-ACP thioesterase mutagenesis studies as well as from structural analysis of the UcFatB binding pocket [17]. Of these 43 positions, 15 were selected based on a previous study which successfully converted a long-chain acyl-ACP thioesterase from *Cuphea viscosissima* to have short-chain preference by interchanging the residues which varied in a sequence alignment [14]. An additional 3 positions were selected based on an early mutagenesis study of the acyl-ACP thioesterase from *Umbellularia californica*, which converted the predominantly C12-specific thioesterase to a predominantly C14-specific thioesterase using sequence analysis of the homolog from *Cinnamomum camphora* to guide the design [7]. Inspection of the *Umbellularia californica* structure, namely residues within the *α1* helix and the *β2*, *β4*, and *β5* helices, led to the identification of 5 additional positions [17]. The remaining positions not encompassed within the criteria above were selected based on a sequence alignment among 6 thioesterases from 3 plant species. Each of the 3 plant species had a representative of a short-chain and a long-chain acyl-ACP thioesterase. The three species included were *Umbellularia californica* (accession numbers Q41635.1 and Q41634), *Cuphea palustris* (accession numbers Q39554 and Q39555), *Cuphea hookeriana* (accession numbers AAC49269.1 and AAC48990.1), and *Cuphea viscosissima* (accession numbers AEM72522.1 and AEM72523.1).

[1] W. Rupilius and S. Ahmad, “Palm oil and palm kernel oil as raw materials for basic oleochemicals and biodiesel,” *Eur. J. Lipid Sci. Technol.*, vol. 109, no. 4, pp. 433–439, 2007.

[2] D. S. Wilcove and L. Pin, “Addressing the threats to biodiversity from oil-palm agriculture,” *Biodivers. Conserv.*, vol. 19, no. 4, pp. 999–1007, 2010.

[3] K. Noweck and H. Ridder, “Fatty Alcohols - Industrial Production,” in *Ullmann’s encyclopedia of industrial chemistry*, 5th ed., Wiley-VCH, 1988, pp. 277–295.

[4] N. J. Hernández Lozada *et al.*, “Highly Active C 8 -Acyl-ACP Thioesterase Variant Isolated by a Synthetic Selection Strategy,” *ACS Synth. Biol.*, vol. 7, no. 9, pp. 2205–2215, 2018.

[5] N. J. Hernández Lozada, T. R. Simmons, K. Xu, M. A. Jindra, and B. F. Pfleger, “Production of 1-octanol in Escherichia coli by a high flux thioesterase route,” *Metab. Eng.*, vol. 61, no. April, pp. 352–359, 2020.

[6] M. J. Grisewood *et al.*, “Computational Redesign of Acyl-ACP Thioesterase with Improved Selectivity toward Medium-Chain-Length Fatty Acids.”

[7] L. Yuan, T. A. Voelker, and D. J. Hawkins, “Modification of the substrate specificity of an acyl-acyl carrier protein thioesterase by protein engineering,” *PNAS*, vol. 92, no. November, pp. 10639–10643, 1995.

[8] F. Jing *et al.*, “Phylogenetic and experimental characterization of an acyl-ACP thioesterase family reveals significant diversity in enzymatic specificity and activity,” pp. 1–16, 2011.

[9] X. Deng, L. Chen, M. Hei, T. Liu, Y. Feng, and G. Y. Yang, “Structure-guided reshaping of the acyl binding pocket of ‘TesA thioesterase enhances octanoic acid production in E. coli,” *Metab. Eng.*, vol. 61, no. January, pp. 24–32, 2020.

[10] T. A. Voelker and H. M. Davies, “Alteration of the Specificity and Regulation of Fatty Acid Synthesis of Escherichia coli by Expression of a Plant Medium- Chain Acyl-Acyl Carrier Protein Thioesterase,” *J. Bacteriol.*, vol. 176, no. 23, pp. 7320–7327, 1994.

[11] Y. J. Choi and S. Y. Lee, “Microbial production of short-chain alkanes,” *Nature*, vol. 502, no. 7472, pp. 571–574, 2013.

[12] S. Sarria, T. G. Bartholow, A. Verga, M. D. Burkart, and P. Peralta-Yahya, “Matching Protein Interfaces for Improved Medium-Chain Fatty Acid Production,” *ACS Synth. Biol.*, vol. 7, no. 5, pp. 1179–1187, 2018.

[13] P. Gordon Roessler and G. Roy, “ACYL-ACP THOESTERASE GENES AND USES THEREFOR,” 8956834 B2, 2015.

[14] F. Jing, L. Zhao, M. D. Yandeau-Nelson, and B. J. Nikolau, “Two distinct domains contribute to the substrate acyl chain length selectivity of plant acyl-ACP thioesterase,” *Nat. Commun.*, vol. 9, no. 1, p. 860, 2018.

[15] M. Politz, R. Lennen, B. Pfleger, and B. Engineering, “Quantification of Bacterial Fatty Acids by Extraction and Methylation,” *Bio Protoc.*, vol. 3, no. 21, 2016.

[16] G. X. Yu, B. H. Park, P. Chandramohan, R. Munavalli, A. Geist, and N. F. Samatova, “In silico discovery of enzyme-substrate specificity-determining residue clusters,” *J. Mol. Biol.*, vol. 352, no. 5, pp. 1105–1117, 2005.

[17] Y. Feng, Y. Wang, J. Liu, Y. Liu, X. Cao, and S. Xue, “Structural Insight into Acyl-ACP Thioesterase toward Substrate Specificity Design,” *ACS Chem. Biol.*, vol. 12, no. 11, pp. 2830–2836, 2017.