



Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily

Serina L. Robinson ^{1,2,3,*}, Megan D. Smith ^{2,3}, Jack E. Richman³, Kelly G. Aukema³, and Lawrence P. Wackett^{3,*}

¹Graduate Program in Bioinformatics and Computational Biology, University of Minnesota, 111 S. Broadway, Suite 300, Rochester, MN 55904, USA, ²Graduate Program in Microbiology, Immunology, and Cancer Biology, University of Minnesota, 689 23rd Ave SE, Minneapolis, MN 55455, USA and ³BioTechnology Institute, University of Minnesota, 1479 Gortner Avenue, Saint Paul, MN 55108, USA

*Corresponding authors: E-mails: robi0916@umn.edu and wacke003@umn.edu

Abstract

Enzymes in the thiolase superfamily catalyze carbon–carbon bond formation for the biosynthesis of polyhydroxyalkanoate storage molecules, membrane lipids and bioactive secondary metabolites. Natural and engineered thiolases have applications in synthetic biology for the production of high-value compounds, including personal care products and therapeutics. A fundamental understanding of thiolase substrate specificity is lacking, particularly within the OleA protein family. The ability to predict substrates from sequence would advance (meta)genome mining efforts to identify active thiolases for the production of desired metabolites. To gain a deeper understanding of substrate scope within the OleA family, we measured the activity of 73 diverse bacterial thiolases with a library of 15 *p*-nitrophenyl ester substrates to build a training set of 1095 unique enzyme–substrate pairs. We then used machine learning to predict thiolase substrate specificity from physico-chemical and structural features. The area under the receiver operating characteristic curve was 0.89 for random forest classification of enzyme activity, and our regression model had a test set root mean square error of 0.22 ($R^2 = 0.75$) to quantitatively predict enzyme activity levels. Substrate aromaticity, oxygen content and molecular connectivity were the strongest predictors of enzyme–substrate pairing. Key amino acid residues A173, I284, V287, T292 and I316 in the *Xanthomonas campestris* OleA crystal structure lining the substrate binding pockets were important for thiolase substrate specificity and are attractive targets for future protein engineering studies. The predictive framework described here is generalizable and demonstrates how machine learning can be used to quantitatively understand and predict enzyme substrate specificity.

Key words: thiolase; *p*-nitrophenyl esters; substrate specificity; machine learning; enzyme activity screen.

1. Introduction

Metabolic pathways for the β -oxidation of fatty acids and production of polyketides, surfactants, β -lactone natural products

and hydrocarbons are initiated by enzymes in the thiolase superfamily (1–4). Carbon–carbon bond formation from the Claisen condensation of two activated fatty-acyl substrates by enzymes in the OleA family of thiolases (3, 5) represents the

Submitted: 20 January 2020; Received (in revised form): 19 May 2020; Accepted: 19 May 2020

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

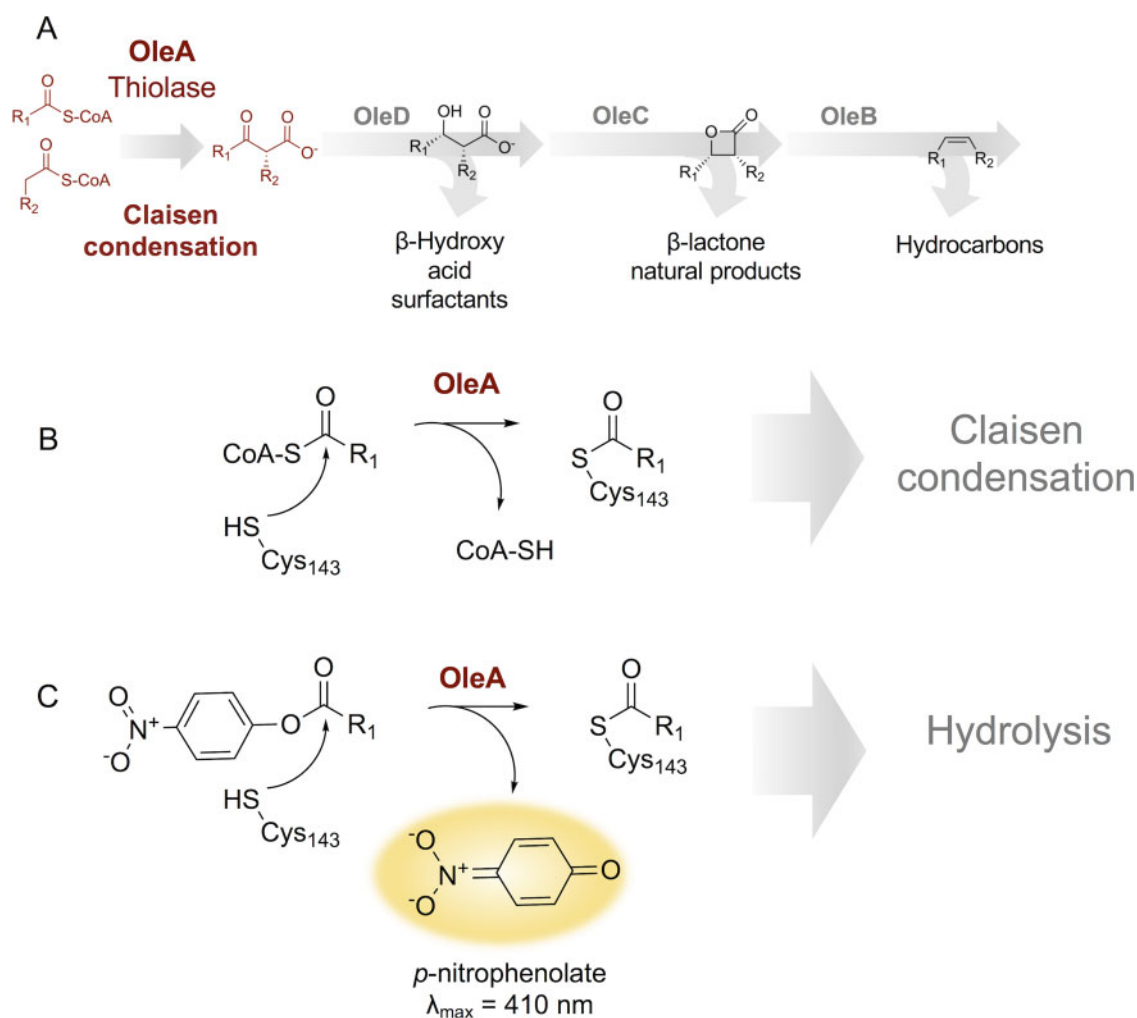


Figure 1. (A) Thiolase enzymes in the OleA family catalyze head-to-head Claisen condensation of two acyl-CoA substrates (maroon) as the first committed step in production of value-added metabolites such as surfactants, pharmaceuticals and hydrocarbons. R_1 , R_2 in characterized OleABCD pathways (3): C_8 – C_{16} (B) Acyl-CoAs as native substrates for OleA-type thiolases acylate the active site cysteine (Cys 143 in *X. campestris*) prior to carbon–carbon bond formation via a Claisen condensation and substrate release. (C) OleA reacts with various *p*-nitrophenyl esters as substrate mimics to produce a *p*-nitrophenolate chromophore that absorbs in the visible range at pH 8.0, providing a rapid readout for enzyme activity.

first committed step in production of the backbone of many value-added bacterial metabolites (Figure 1A). Previously, we demonstrated that swapping OleA from *Shewanella oneidensis* with *Stenotrophomonas maltophilia* OleA altered the chain length and profile of hydrocarbons produced downstream by OleBCD enzymes (3). Within the broader thiolase superfamily, Prather and colleagues used a rational design approach to alter thiolase substrate specificity for metabolic engineering of the reverse β -oxidation pathway (6). Thiolase substrate specificity is particularly critical for metabolic engineering applications since it often ‘sets’ the chemical composition of downstream products.

Thiolases typically use a ping-pong mechanism whereby an activated substrate acylates the active site cysteine and remains tethered covalently until the second substrate binds and the acyl group is transferred, resulting in carbon–carbon bond formation. The majority of well-characterized thiolases are FabH-type enzymes with a single, deep hydrophobic substrate channel (7). FabH initiates the two-carbon elongation of fatty acids iteratively by condensing a malonyl unit onto a growing backbone with the concurrent release of CO_2 during each

catalytic cycle. In contrast, OleA-type thiolases have two hydrophobic substrate channels instead of one and catalyze the non-decarboxylative Claisen condensation of two long-chain fatty acids (8, 9). At present, the only biochemically and structurally characterized OleA is from *Xanthomonas campestris*, a bacterial plant pathogen (5, 8, 9).

OleA enzymes characterized to date accept fatty acid substrates activated with coenzyme A (CoA), which are costly feedstocks for biotechnological applications. During the first step of the OleA catalytic cycle, acyl-CoA substrates undergo transesterification to the active site cysteine (Figure 1B) prior to carbon–carbon bond formation (5, 8, 9). Recently, we discovered that OleA-family enzymes also hydrolyze *p*-nitrophenyl esters (pNPs) to release *p*-nitrophenolate as a rapid colorimetric readout for OleA activity (Figure 1C). Here, we used this assay to screen 15 different pNP substrates against a library of 73 OleA sequences from taxonomically diverse bacteria sharing as low as 13.8% pairwise amino acid identity (Supplementary Table S1). We then trained machine learning models on our paired enzyme–substrate dataset to quantitatively predict thiolase activity with different pNPs.

Machine learning is gaining traction in chemical biology as a powerful technique for the prediction of enzyme substrate specificities (10). Support vector machines and ensemble learning methods achieved high accuracy for the prediction of amino acid substrates for nonribosomal peptide synthetase adenylation domains (11, 12). Integration of these machine learning algorithms within a larger predictive pipeline known as antiSMASH has improved structural prediction of natural products from genomic information (13). Machine learning-based methods have also been applied to predict substrate specificities of other protein families including glycosyltransferases (14), acyl-CoA ligases (15) and proteases (16–18).

While thiolases have many applications in synthetic biology, quantitative insights into how physicochemical properties of residues in the substrate binding pockets affect substrate specificities are lacking. Using thiolases as biological catalysts to produce desired compounds requires a deeper understanding of their natural substrate range. Accurate prediction of substrate specificity will aid in expanding the toolbox of standardized parts for carbon–carbon bond formation and enable enzymatic production of compounds with backbones of desired chain length and composition. The experimental and machine learning frameworks described here may also be generalized to learn the substrate specificity rules of different enzymes classes. This approach can be integrated into an automated workflow for machine learning-guided parts selection for custom metabolite production and other applications in synthetic biology.

2. Materials and methods

2.1. Chemicals and reagents

Chemical syntheses of nine pNPs from their corresponding carboxylic acid precursors was carried out using the method of Engström et al. (19). Detailed synthetic procedures are in the [Supplementary Methods](#), and proton nuclear magnetic resonance ($^1\text{H-NMR}$) spectra of all synthesized compounds are in [Supplementary Figure S1](#). Purchased carboxylic acids from Sigma-Aldrich were 3-cyclopentylpropanoic acid, 7-phenylheptanoic acid, 3-(4-chlorophenoxy)propanoic acid, 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoic acid, 2-(2-butoxyethoxy)acetic acid and 6-heptynoic acid. 2,2-Dimethylhexanoic acid and heptanoic acid were purchased from TCI-America, and 6-azidoheptanoic acid was purchased from Cayman Chemical Company. Additionally, five commercially available pNPs were obtained from the companies indicated: *p*-nitrophenyl trimethylacetic acid, *p*-nitrophenyl decanoic acid and *p*-nitrophenyl dodecanoic acid (Sigma-Aldrich), *p*-nitrophenyl biotin and *p*-nitrophenyl benzoic acid (Alfa Aesar) and *p*-nitrophenyl hexanoic acid (TCI-America).

2.2. Strain and plasmid construction

Synthetic DNA for a library of 73 OleA homologs were cloned with N-terminal 6 \times -His tags into pET28b+ vectors at NdeI and XhoI restriction sites and transformed into T7 Express Competent *Escherichia coli* cells (NEB C2566I) by the U.S. Department of Energy Joint Genome Institute. Accession numbers ([Supplementary Table S2](#)) and codon-optimized plasmid sequences ([Supplementary Material S1](#)) are available. Bioinformatic methods for gene selection and construct design are described in detail by Smith et al. (20).

2.3. Culture conditions and whole-cell assays

Cells were grown at 37°C shaking at 250 rpm in 5 ml of lysogeny broth in test tubes to an optical density of 0.3 and induced with a final concentration of 1 mM isopropyl β -D-1-thiogalactopyranoside. Induced cultures were incubated for 40–43 h at 16°C with agitation at 250 rpm. Substrate specificity screening was conducted using the whole-cell thiolase assay protocol described by Smith et al. (20). Briefly, cells were normalized to an optical density of 0.1 per 200 μL of cells in 50 mM Tris–HCl (pH 8.0) buffer and incubated for 2 h with 63 μM polymyxin B sulfate to render them porous to small molecules. Cells expressing each of the 73 enzymes were tested with 15 different pNP substrates added at a maximum final concentration of 200 μM . Poor solubilities of some substrates resulted in minor precipitation. In a previous study, it was demonstrated that adding substrates above their limit of solubility did not significantly interfere with assay readout (20). Absorbance was read in 96-well clear bottom plates (Genesee Scientific) every minute for 60 min at 410 nm and 37°C using a SpectraMax Plus 384 microplate reader (Molecular Devices). All 73 enzymes were tested with 15 pNP substrates, first with technical duplicates from the same induction batch, and then in biological triplicates with three independently grown cultures for all of the active enzyme–substrate pairs.

2.4. Data preprocessing

Absorbance values were normalized by subtracting averaged values from triplicate empty pET-28b+ vector controls on each plate. Absorbance values for each plate were converted to nmol *p*-nitrophenolate using *p*-nitrophenolate standard curves run in parallel on the same plate. Activity was calculated using a ‘rolling window’ linear regression method by calculating slopes for all overlapping 15 min intervals over the course of the first 45 min of each reaction, with the exception of *p*-nitrophenyl 2-(2-butoxyethoxy)acetate which hydrolyzed rapidly in buffer such that catalyzed reaction rates above background could only be measured for 5 min before absorbance reached the maximum detection limit. The maximum slope with an $R^2 \geq 0.9$ was converted to enzyme activity (nmol *p*-nitrophenolate/OD 1.0/h). Activities for each enzyme–substrate pair were averaged across biological triplicates and are reported in [Supplementary Table S2](#).

2.5. Physicochemical feature engineering

For each of the 15 pNP substrates, 153 chemical properties were calculated using the RcpI and ChemmineR packages in R (21, 22). Since many of the chemical properties were correlated, we performed dimensionality reduction using principal component analysis. Loadings of the first seven principal components included in analysis are detailed in [Supplementary Figure S2](#). To extract protein sequence features, we used DECIPHER, a structure-based aligner (23) that uses local sequence context to align each of the 73 OleA sequences with the crystal structure of *X. campestris* (PDB ID: 4KU5). We extracted spheres of residues from each aligned protein with radii 8, 10, 12 and 14 Å from the α -carbon of the active site cysteine. Twelve angstrom was selected as the best sphere size for model training because it was the smallest radius that encompassed all residues lining both substrate binding pockets, thereby keeping the number of features relative to total training data points reasonable. Each amino acid in the 12 Å radius was encoded as a vector of principal components of its physicochemical properties as described

by Atchley et al. (24). The codon diversity index calculated by Atchley et al. was not included analysis, therefore each of the 84 amino acids within a 12 Å sphere were encoded by four indices corresponding to polarity, molecular volume, secondary structure and electrostatic charge. Global protein properties used as features including instability index, isoelectric point, molecular weight and Kyte-Doolittle hydrophobicity index (see [Supplementary Table S3B](#) for full set of indices) were calculated using the *Peptides* package in R (25).

2.6. Machine learning

Protein and chemical features were concatenated into a single numeric vector describing each enzyme–substrate pair ($n = 1095$). Features with near-zero variance were removed using the *nearZeroVar* function from the *caret* package in R (26). The data were split randomly with stratified sampling by activity to achieve roughly equal proportions of active and inactive enzymes in 75% training and 25% testing sets. R version 3.6.1 and *caret* were used to evaluate all models (26). Grid search was used to tune model hyperparameters by 10-fold cross validation repeated in triplicate. For the random forest algorithm, model hyperparameters that were tuned included the number of variables randomly sampled as candidates at each split, the minimum size of terminal nodes and the splitting rule methods. All forests were grown to a size of 1000 trees and the permutation method was used to compute relative feature importance as implemented in the *ranger* package in R (27). A description of hyperparameters tuned for other machine learning models tested were detailed in [Supplementary Figure S3](#). The distribution of training and testing set prediction performances were examined by 1000 independent, random training-test splits ([Supplementary Figure S4](#)).

Three different machine learning algorithms were evaluated for classification of enzyme–substrate pairs: random forest, naïve Bayes and feedforward neural networks. Receiver operating characteristic curves and confusion matrices were used to compare model performances. For the 550 enzymes classified as active, we further trained regression models to quantitatively predict enzyme activity. Enzyme activity values displayed a right-skewed distribution so a \log_{10} transformation was applied, resulting in an approximately normal distribution. Three different machine learning algorithms were evaluated for regression: random forest, elastic net and multivariate adaptive regression splines. Root mean square error (RMSE) and R^2 values were used to assess performance. Models were further evaluated using leave-one-compound-out and leave-one-taxon-out validation ([Supplementary Figure S5](#)).

2.7. Homology modeling, phylogenetics and bioinformatics

An approximate maximum-likelihood phylogeny for the OleA amino acid sequences was estimated using FastTree version 2.1 (28) using the Jones–Taylor–Thornton model and assuming a single rate of evolution for each site known as the “CAT” approximation. Between-groups analysis was used to detect differential residues between broad and narrow specificity enzymes using the *bgafun* package in R (29). Homology models for each of the 73 OleA proteins were built using the Phyre2 server (30). Solvent-accessible surface area and cavity volumes for each of the homology models were calculated using CastP version 3.0 with a 2.2 Å radius probe (31).

2.8. Data and code availability

Raw data and scripts to reproduce analyses, figures and tables are available at <https://github.com/serina-robinson/thiolase-machine-learning/>. An interactive web application with a searchable database and predictive models trained on the complete dataset are also available at z.umn.edu/thiolases (shortened URL) and srobinson.shinyapps.io/thiolases (permanent URL). The DNA constructs, provided in [Supplementary Material S1](#), will be provided upon request. The DNA constructs were provided by the United States Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02-05CH11231. DNA requests will be honored with the completion of a Materials Transfer Agreement as required by our contracts with the U.S. Department of Energy.

3. Results

3.1. Selection and screening of chemically diverse pNP substrates

We aimed to construct the first quantitative map of natural variation in substrate specificity for OleA-type thiolases. Recently, we developed a whole-cell assay using pNPs to rapidly screen thiolase enzymes without time-consuming protein purification, quenching and extraction steps (20). Here, we used this assay to screen a library of thiolase enzymes with 15 diverse pNP compounds. To select candidates for screening, we web scraped Sigma-Aldrich pages to identify commercially available carboxylic acids ($n = 3572$). Based on clustering analysis using the Tanimoto coefficient ([Supplementary Figure S6A](#)), we selected 15 pNPs that spanned a wide range of chain lengths, heteroatomic composition and functional groups ([Supplementary Figure S6B](#)). The final library of 15 pNPs were synthesized from their carboxylic acid precursors as described in the [Supplementary Methods](#) or purchased directly as described above.

We screened each of the 15 pNPs against a panel of 73 bacterial thiolase enzymes heterologously expressed in *E. coli* for a total of 1095 enzyme–substrate pairs ([Figure 2](#)). All enzymes were active with at least two pNP substrates and, with the exception of *p*-nitrophenyl benzoate, all substrates reacted with at least one enzyme. The most broadly reactive pNP was *p*-nitrophenyl 2-(2-butoxyethoxyacetate) for which 71 out of 73 enzymes yielded product. The enzyme with the highest average activity was natively from *Kytococcus sedentarius*, a common constituent of the human skin microbiome. The *K. sedentarius* thiolase reacted with 14 different pNPs and was among the top 3 most active enzymes for 10 of these substrates.

3.2. Broad substrate specificity among actinobacterial and gammaproteobacterial OleA clades

We first examined whether OleA thiolase substrate specificity had a phylogenetic signal. Our enzyme library included sequences from 7 different phyla and 68 different bacterial genera (20). We observed three monophyletic clades that exhibited a high level of activity across a wide range of pNP substrates ([Figure 2](#)). Clade I contained gammaproteobacterial thiolases within the *Xanthomonadaceae* including *Chromatococcus*, *Luteimonas*, *Thermomonas* and the structurally characterized *X. campestris* OleA (8, 9). The other two highly active clades (II and III) consisted of thiolases from Actinobacteria including *Kytococcus*, *Mobilicoccus*, *Dermatophilus* and *Kocuria*. Even within these clades, pNP substrate profiles were variable across enzymes from



Figure 2. Approximate maximum-likelihood phylogeny of 73 OleA protein sequences paired with heatmap of enzyme activities across 15 different pNP substrates. †Enzyme activities were measured as the log₁₀ of nmol p-nitrophenolate produced over the course of one hour by an *E. coli* BL21 culture with an OD of 1.0 per 200 µL of cells heterologously expressing OleA. *Average enzyme activity is across three biological replicates for each substrate screened. Circle sizes are scaled to average activity for each enzyme across all substrates. Clade I (orange shading) represents the most active clade of Gammaproteobacterial enzymes including the *X. campestris* OleA designated with a gray square for which the crystal structure (PDB ID: 4KU5) has been solved (8, 9). Clades II and III (blue shading) represent the most active clades of OleA enzymes found in Actinobacteria.

closely related organisms and activity could not be predicted on the basis of phylogeny alone.

3.3. Relationship between binding pocket volume and bulky substrate preferences

Actinoplanes atraurantiacus had the highest preference for *p*-nitrophenyl trimethylacetate across all enzymes in the library, even outperforming *K. sedentarius* and other highly active thiolases. *Brachybacterium paraconglomeratum* and *Halobacteriovorax marinus* also had thiolases with 'bulky' substrate specificity, exhibiting higher activity with *p*-nitrophenyl 2,2-dimethylhexanoate than with the C₆–C₇ length substrates preferred by the majority of enzymes screened. Since the *tert*-alkyl substituents of the trimethylacetate and 2,2-dimethylhexanoate compounds

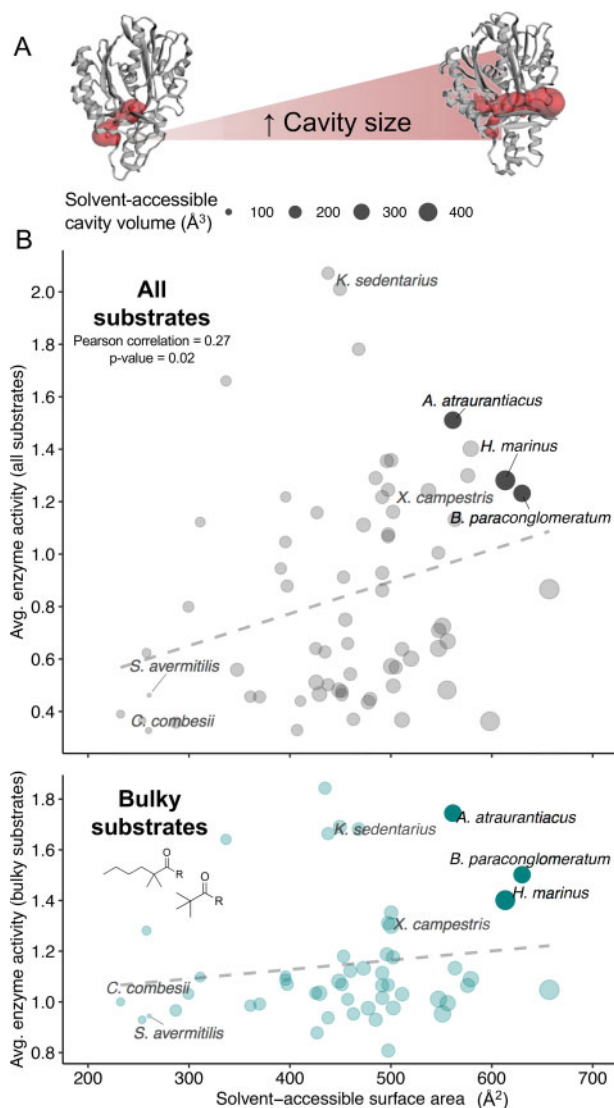


Figure 3. (A) Predicted solvent-accessible cavity volumes of the 73 OleA enzymes analyzed in this study ranged from 95.3 to 478.9 Å³. (B) Relationship between the solvent-accessible surface area with the average enzymatic activity across all substrates (black) and 'bulky' substrates with *tert*-alkyl-substituted α -carbons (teal, *p*-nitrophenyl trimethylacetate and *p*-nitrophenyl 2,2-dimethylhexanoate). Enzymes from *Actinoplanes atraurantiacus*, *Brachybacterium paraconglomeratum* and *Halobacteriovorax marinus* have among the highest calculated cavity volumes and the highest preferences for substrates with *tert*-alkyl-substituted α -carbons.

fill a larger volume near the α -carbon than most of the other pNPs tested, we hypothesized that the volume of the substrate binding pocket might affect enzyme activity. We constructed homology models for the 73 enzymes and used a computational geometry method (31) to calculate the solvent-accessible surface area and binding pocket volumes (Figure 3). We observed a weak positive association between solvent-accessible surface area and average enzyme activity (Pearson correlation = 0.27, P-value = 0.02). Three enzymes with unusually high preferences for bulky α -carbon substrates (*Actinoplanes*, *Halobacteriovorax* and *Brachybacterium*) also had among the highest predicted solvent-accessible surface areas (Figure 3). However, we noted other enzymes with lower predicted solvent-accessible surface areas also displayed comparable activity with bulky substrates while some enzymes with large predicted surface areas did not display a preference for bulky substrates (Figure 3). Although care was taken to thread all models to the same template to avoid template bias, results must be interpreted with caution since homology modeling does not recapitulate structural dynamics and variations in pocket volume during substrate binding.

3.4. Statistical identification of specificity-determining residues T292 and L203

Approximately one-third of the enzymes we tested were active with 10 or more pNP substrates, suggesting that a large number of OleA-type thiolases had a broad substrate range, including the highly active enzyme from *K. sedentarius*. In contrast, enzymes that reacted with five substrates or fewer (Supplementary Table S2) were defined to have narrow substrate specificity. We used a multivariate statistical method termed between-groups analysis (BGA) to detect residues that differed between broad- and narrow-specificity thiolases (29). BGA consistently identified two key residues lining substrate binding channels in the crystal structure that were conserved in more than 50% of enzymes in each group. The identification of these two residues was robust to different splits and group sizes for broad and narrow specificity enzymes. Residue 292 was located towards the end of the substrate channel A in the *X. campestris* crystal structure. It was a conserved Thr in broad specificity enzymes that was absent in enzymes with narrow specificity (P-value < 0.05, Fisher exact test). All but one of the top 30 most active enzymes in our dataset had a Thr aligning with position 292. Narrow specificity thiolases had variety of charged (Asp, Glu, Arg) or aliphatic (Ile, Leu, Val) residues aligning with position 292 instead. The second key residue aligned with residue 203 located at the end of substrate channel B in the *X. campestris* structure. This residue was a conserved Leu/Ile in the top 32 broad specificity enzymes and a Gly or Val among narrow specificity enzymes that accepted four or fewer substrates (P-value < 0.05, Fisher exact test).

3.5. Machine learning prediction of paired enzyme–substrate activity relationships

While BGA was useful to identify conserved amino acids between enzyme groups that may affect substrate specificity, it could not fully capture the complexity of enzyme–substrate relationships; this task is better suited for machine learning. We next evaluated the performance of different machine learning algorithms to predict substrate specificity from a combination of physicochemical protein and substrate features.

To construct a set of substrate features, we calculated 153 chemical descriptors and used principal component analysis to reduce these descriptors into linearly uncorrelated principal components (PCs, [Supplementary Figure S2A](#)). The first seven chemical PCs were able to explain 92% of the variance between substrates ([Supplementary Figure S2B](#)). We extracted the absolute values of the PC loadings to determine the overall contribution of different chemical descriptors to the PCs ([Supplementary Figure S2C](#)). We observed that the top seven PCs corresponded broadly to the following chemical properties: molecular weight (PC1), molecular connectivity (PC2), aromaticity (PC3), solubility (PC4), oxygen content (PC5), nitrogen content (PC6) and chlorine content (PC7).

A structure-based sequence alignment method was used to align each protein sequence in the training set with the crystal structure of an OleA-type thiolase with substrates bound (PDB ID: 4KU5). All residues that aligned within 12 Å of the active site cysteine (Cys 143) were encoded into a numerical vector of physicochemical indices corresponding to polarity, molecular volume, secondary structure and electrostatic charge. We found this method of amino acid featurization increased training set classification accuracy 2% over a one-hot encoding method which does not capture information about the physicochemical properties of amino acids. We also calculated macromolecular protein properties based on the full-length input sequences including hydrophobicity, isoelectric point, molecular weight and instability indices ([Supplementary Table S3B](#)). All chemical and protein features were then concatenated into a single vector representing the unique physicochemical signature of each enzyme–substrate pair.

3.6. Substrate aromaticity and molecular connectivity influences enzyme–substrate pairing

Of the 1095 enzyme–substrate pairs tested experimentally, 550 were active and 545 were inactive ([Figure 4A](#)). We first evaluated three different machine learning algorithms for binary classification of enzyme–substrate pairs as active or inactive. Of the three algorithms tested (random forest, naïve Bayes and feed-forward neural networks), we observed the highest classification accuracy (area under the receiver operating characteristic curve = 0.89) with the random forest model ([Table 1](#), [Supplementary Figure S3A](#)). From 1000 random training-testing dataset splits we obtained an average testing set classification accuracy of $81.9 \pm 2.2\%$ ([Supplementary Figure S3A](#)). Our model indicated that chemical features were universally more important than sequence features for classification of enzyme–substrate pairs as active or inactive ([Figure 4B](#)). In particular, the substrate aromaticity index (PC3) was the most important feature followed by the molecular connectivity index (PC2) including valence chi chain descriptors, Kier molecular shape indices and number of rotatable bonds ([Supplementary Figure S2](#), [Supplementary Table S3](#)). Activity increased with carbon chain length to a certain point (7–8 carbons) beyond which activity decreased, as observed with *p*-nitrophenyl decanoate and *p*-nitrophenyl dodecanoate. Overall, the majority of enzymes tested preferred substrates with a higher number of rotatable C–C or C–O bonds and a chain length of 6–7 carbons instead of longer or shorter chain lengths tested here.

Notably, our machine learning algorithm identified channel residue 292 to be an important sequence feature for classification accuracy ([Figure 4C](#)). This residue had also been identified previously by BGA, further supporting a Thr in position 292 is likely important for substrate specificity. Moreover, a number of

hydrophobic residues lining both binding pockets were identified by our model to be important including V287 and A173. We postulate these may assist in maintaining a hydrophobic environment within the binding pockets as was suggested in earlier crystallographic studies (8, 9). In *X. campestris*, an analysis of the residues lining substrate channels revealed the most common residues are Val, Leu and Ile (8). These aliphatic side chains promote binding of long-chain fatty acid substrates and likely play a role in determining which *p*NP substrates will bind and react in the OleA binding pocket.

Among the full-length protein indices included in our model, two protein indices for helix/turn propensity, Kidera and FASGAI (32, 33), were also important for prediction accuracy. We observed a negative correlation between the Kidera factor for helix/turn propensity of an enzyme and its average activity across all substrates (Pearson correlation = -0.23 , P -value = 0.05). Overall, enzymes with lower helix propensity tended to be more active across all substrates, suggesting secondary structure flexibility enhances broad substrate specificity. This is consistent with the importance of Val and Ile residues in our models since the flexibility of their aliphatic side chains likely allows a wider variety of substrates to bind.

3.7. ‘Pinch point’ residues and substrate oxygen content influence enzyme–substrate regression models

We next examined whether we could quantitatively predict enzyme activity using regression models trained on the 550 active enzyme–substrate pairs. We evaluated three different machine learning algorithms (random forest, elastic net and multivariate adaptive regression splines). Again, random forest outperformed other models with a testing set R^2 of 0.75 ([Table 2](#)) and a testing RMSE of 0.243 ± 0.016 estimated from 1000 random training-testing dataset splits ([Supplementary Figure S3B](#)).

Our regression models showed that PC5 (oxygen content) was the most important chemical feature for quantitative prediction of enzyme activity ([Figure 4B](#)). This is consistent with our findings that the *p*-nitrophenyl 2-(2-butoxyethoxy)acetate substrate had the highest average activity across all substrates and also has the highest number of oxygen atoms. PCs corresponding to molecular connectivity and solubility were the second and third most important features, respectively. The molecular connectivity index (PC2) was positively correlated with average enzyme activity (Pearson correlation = 0.46 , P -value < $2.2e^{-16}$). An examination of the loadings of PC2 further revealed positive associations between the number of atoms in the largest chain, the number of rotatable bonds and enzyme activity. These results further support the hypothesis for OleA-type thiolases preferring ‘spacer’ methylene carbons between the α -carbon and other functional groups such as phenyl, cyclic aliphatic, alkynyl or azido groups.

Residues 172, 173, 284, 287 and 316 were also identified as important by our regression algorithm and are known residues lining the substrate binding channels in the *X. campestris* structure ([Figure 4C](#)). Val 287 in the *X. campestris* structure plays a role in directing the path of the alkyl chain of the native substrate to curve around the hydrophobic Ile 345 side chain (8). Ile 284 is also in a critical position and likely interacts with Thr 292 based on the *X. campestris* crystal structure. A triad of residues (I284, A261 and T292) in channel A are predicted to form a ‘pinch point’ to mediate chain length specificity (9). Our results support this hypothesis and suggest these ‘pinch point’ residues are prime targets for rational design studies to alter thiolase substrate specificity.

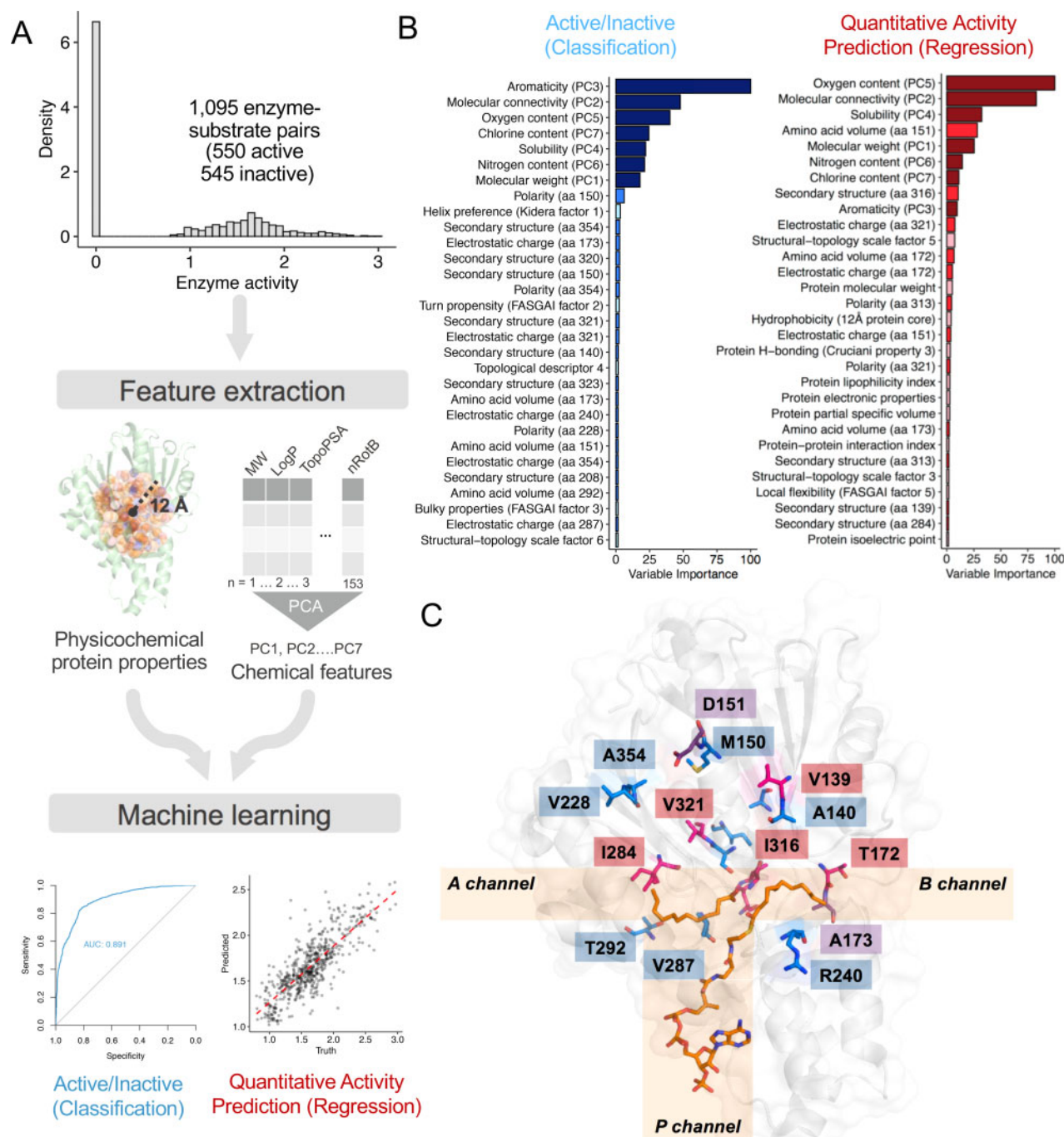


Figure 4. (A) Machine learning workflow used in this study. AUC, area under the receiver operating characteristic curve. (B) Variable importance scores for classification and regression models. (C) Important residues mapped onto the *X. campestris* structure (PDB ID: 4KU5) with fatty acid and acyl-CoA substrates bound. Residue colors correspond to variable importance in the classification model (blue), regression model (red) or both models (purple).

Table 1. Machine learning classification results

Machine learning algorithm	Training classification accuracy	Testing classification accuracy	Testing 95% confidence interval
Random forest	0.826	0.839	0.789–0.880
Feedforward neural network	0.732	0.777	0.722–0.826
Naïve Bayes	0.586	0.645	0.585–0.701

See [Supplementary Figure S3A](#) for extended results.

Table 2. Machine learning regression results

Machine learning algorithm	Training RMSE	Training R ²	Testing RMSE	Testing R ²
Random forest	0.254	0.625	0.219	0.745
Multivariate adaptive regression splines	0.276	0.557	0.252	0.642
Elastic net	0.275	0.549	0.278	0.564

See [Supplementary Figure S3B](#) for extended results.

RMSE = Root mean square error.

3.8. Leave-one-out validation

To assess biases in our training set and determine how our models would perform with pNP substrates not included in model training, we performed leave-one-compound-out validation. We trained 15 separate models by omitting one compound completely from each model during training and then evaluating prediction accuracy for the held-out compound. Results from leave-one-compound-out validation for both classification and regression algorithms revealed that our model performed best with long-chain alkyl substrates including *p*-nitrophenyl dodecanoate and *p*-nitrophenyl 7-phenylheptanoate ([Supplementary Figure S5](#)). Accuracy decreased for more chemically distinct and polar substrates such as *p*-nitrophenyl azidohexanoate and *p*-nitrophenyl 2-(2-butoxyethoxy)acetate. Overall, the leave-one-compound-out analysis indicated our models performed relatively well (>74% classification accuracy) for most alkyl pNP substrates and underperformed with pNP compounds containing chemical moieties not included in the training set such as azides, ethers and trimethyl groups.

To test the phylogenetic bias of our model, we performed leave-one-taxon-out validation. OleAs in the gene library had been sampled from organisms belonging to 10 different taxonomic classes ([Figure 2](#)). We trained 10 separate models by withholding all sequences from organisms belonging to one taxonomic class and then testing prediction performance on the omitted class ([Supplementary Figure S5](#)). We found that both classification and regression models performed more robustly with higher accuracy and RMSE values with leave-one-taxon-out validation than the leave-one-compound-out performance. This reflects that there is much more variability in average activity across different pNP compounds than across thiolases from different taxonomic classes. The two taxonomic classes which most significantly decreased classification and regression model performance when omitted were Gammaproteobacteria and Actinobacteria. This is consistent with the largest proportion of OleAs in the training set belonging to organisms from these classes.

3.9 Predictive web application

To make our machine learning models accessible to users with all levels of computational expertise, we created an interactive web application (z.umn.edu/thiolases). This web interface allows users to upload protein or nucleotide FASTA files of OleA-type thiolase sequences and uses machine learning models trained on the entire dataset to make rapid predictions for pNP substrate specificity. It provides predictions for activity with each of the 15 pNP substrates tested in this study as well as probability scores ranging from 0.5 (low confidence) to 1 (high confidence). Chemical structures of predicted substrates are displayed through the web interface ([Supplementary Figure S7](#)) and results are downloadable in tabular format for further analysis.

4. Discussion

Over half of the thiolases tested were active with at least half of the pNPs tested. The remarkably broad substrate specificity in a large fraction of OleA-type enzymes in our library corresponds well with a previous study where hydrocarbons were extracted from twelve bacterial strains containing the complete *ole*ABCD cluster (3). The authors reported that some bacteria, including strains from the genera *Xanthomonas* (Clade I) and *Kocuria* (Clade III), had *ole*ABCD pathways that produced up to 15 different hydrocarbons identifiable by GC-MS (3). In our study, enzymes from *Xanthomonas* and *Kocuria* genera reacted with 11–14 different pNP substrates out of 15 tested. It is likely we have only sampled the ‘tip of the iceberg’ in terms of the number of potential substrates accepted by some highly active enzymes in this study such as the OleA from *K. sedentarius*.

In contrast, some enzymes tested such as the OleAs from *Shewanella putrefaciens* and *Psychromonas aquimarina* had narrow substrate specificity and weak activity. One explanation for the low activity of OleA enzymes from *Shewanella* and *Psychromonas* are their native preferences for a specific class of polyunsaturated fatty acid (Pfa) precursors that are produced by polyketide/fatty acid synthases encoded in a five-gene operon (*pfa*ABCDE) upstream of the *ole*ABCD cluster. Previously, Sukovich and colleagues extracted and verified a single 3,6,9,12,15,19,22,25,28-hentriacontanonaene product from the *Shewanella ole*ABCD pathway (34). The authors postulated this product was derived from OleA-catalyzed condensation of two CoA-activated molecules of hexadeca-4,7,10,13-tetraenoic acid produced by the *pfa* operon. This was further supported by recent work in *Shewanella pealeana* demonstrating that OleA interacts with the Pfa synthase *in vivo* and mediates the transfer of Pfa precursors to the OleBCD complex to facilitate polyunsaturated hydrocarbon biosynthesis (35). We hypothesize here that the narrow substrate specificity observed in OleAs *Shewanella* and *Psychromonas* may be due to the co-evolution of *pfa* and *ole* sequences. To further investigate this, we identified 92 new OleA sequences from organisms that also encode *pfa* genes ([Supplementary Material S2](#)). We used our machine learning models to make predictions for the reactivity of these enzymes with 15 pNP substrates. We found that the predicted number of substrates accepted by OleAs co-localized with *pfa* genes was between 2 and 4 (average of 3), compared to an average of 8 substrates accepted by OleAs from organisms without *pfa* genes ([Supplementary Table S4](#)). These predictions provide preliminary support for the hypothesis that *pfa*-associated OleA co-enzymes may have narrower substrate specificity than non-*pfa*-associated OleAs. Further research is required into how thiolases may have co-evolved with precursor biosynthetic enzymes such as Pfa synthases to affect OleA substrate selectivity.

The only substrate which was not active with any enzyme tested was *p*-nitrophenyl benzoate. The compound had been purchased, and we verified its purity by ¹H-NMR and mass

spectrometry (Supplementary Figure S1). We speculate the complete lack of activity with *p*-nitrophenyl benzoate may be due to the phenyl ring, lacking intervening methylene carbons, being bound in an unproductive manner such that the substrate carbonyl carbon is not accessible to the active site cysteine. Alternatively, steric hindrance at the α -carbon due to the bulky phenyl group could prevent cysteine attack. Other aromatic compounds tested, including *p*-nitrophenyl 3-(4-chlorophenoxy)propanoate, *p*-nitrophenyl 3-(5-phenyl-1,3,4-oxadiazol-2-yl)propanoate and *p*-nitrophenyl 7-phenylheptanoate, were also relatively poor substrates compared to those with higher aliphaticity. In general, our data suggest that placement of aromatic groups further away from the α -carbon results in higher reactivity with OleA-family thiolases.

Based on our results, we infer the optimal pNP substrate for OleA enzymes has a side chain with at least 5–7 rotatable C–O or C–C bonds. Adequate spacing of 5–7 carbons between the α -carbon and other functional moieties such as alkynes, azides or cyclic aliphatic groups correlated with higher activity levels. Increasing the number of rotatable bonds was also positively associated with activity. Interestingly, the number of rotatable bonds in a molecule is considered to be a good descriptor of oral bioavailability of drugs (36). Overall, this work provides a foundation in the natural variation in OleA-family thiolase substrate specificity to support future applications in synthetic biology such as using enzyme cascades to produce drug-like molecules. Recent advances have been made using biocatalytic cascades for the total synthesis of therapeutics such as the HIV drug, *islatravir* (37). We envision an expansion of the experimental and machine learning frameworks described here to aid in the design of enzyme and substrate libraries and ultimately use engineered thiolases for the production of bioactive molecules.

Substrates with terminal azide and alkyne moieties unexpectedly exhibited higher average reactivity with our enzyme library than the ‘unmodified’ alkyl *p*-nitrophenyl hexanoate and *p*-nitrophenyl heptanoate substrates. Substrates with azido- or alkynyl-groups were the second and third most reactive substrates screened, respectively. This represents the first experimental evidence of OleA-type thiolases reacting with substrates with ‘clickable’ functional groups. Copper(I)-catalyzed azide-alkyne cycloadditions, known as ‘click’ chemistry reactions, can be run under mild conditions that are particularly well-suited for biological systems. Recently, a reliable method for the synthesis of azides from primary amines opened new chemical possibilities for screening large libraries of clickable substrates (38). The ability for OleA-family thiolases to accept these compounds expands their applications in biological imaging and drug design through attachment of fluorophores or pharmacophores.

The overall goal of the assay was to provide a rapid screening method to identify thiolases with desired substrate specificity profiles that both express well and are active. One limitation of the whole-cell pNP assay is that the colorimetric readout from whole cells necessarily conflates protein expression with measured enzyme activity levels (20). All thiolases screened in this study were active with at least two pNPs (Supplementary Table S2) suggesting all proteins were expressed. Since levels of protein expression can vary between experiments depending on exact induction time and cell density, we screened each of the active enzyme–substrate pairs in biological triplicates. However, we cannot rule out that the measured weak activity of some enzymes is a result of consistently poor protein expression levels across multiple experiments.

Another limitation of the pNP assay is that it only measures the first step of the mechanism: transesterification. We are currently investigating Claisen condensation assay development since preferred substrates for the second step of the thiolase reaction (condensation) may be different than for transesterification. Previously, we purified several thiolases active with pNPs and measured their condensation activity with a variety of acyl-CoA substrates by GC-MS (20). However, this approach is not feasible for screening large libraries of enzymes. The development of a rapid assay to directly measure condensation activity will expand the applications of thiolases in synthetic biology.

In summary, we have quantitatively mapped the substrate specificity of thiolase enzymes through whole-cell assays, structural homology modeling, binding pocket analysis and machine learning. We surveyed a library of thiolase variants from taxonomically diverse organisms to assess natural variation in substrate scope and made predictions for new enzymes. This dataset will serve as a baseline for protein engineering studies to benchmark the performance of engineered enzymes against natural variants. Future experimental efforts will focus on altering key residues identified through our analysis to engineer thiolases to have high activity and desirable substrate specificity profiles.

The application of machine learning to predict substrate selectivity is generalizable to other enzyme families beyond thiolases and may be further improved through the use of ensemble or deep learning methods. While the feedforward neural network algorithm tested here did not outperform random forest, this is likely a result of our modest dataset size (1095 data points). We anticipate with larger datasets (>10 000 data points) that deep learning will outperform shallow learning algorithms like random forest. Overall, this work is a stepping stone towards a new frontier in biology where the combination of terabytes of meta-omic data with artificial intelligence can be used to learn complex patterns undetectable to the human eye. Here, we demonstrated a proof-of-principle for machine learning using physicochemical features to predict enzyme substrate specificity. Ultimately, we envision experimental and computational methods such as those described here can be combined to advance the design of novel biological parts for the biosynthesis of high-value metabolites.

Supplementary data

Supplementary Data are available at SYN BIO online.

Acknowledgements

We acknowledge Yasuo Yoshikuni, Jan-Fang Cheng and Miranda Harmon-Smith at the U.S. Department of Energy Joint Genome Institute for their support, insights and discussions during construct design. Barbara Terlouw is acknowledged for critical feedback on the manuscript and on machine learning strategies with support from Janani Durairaj. We are also grateful to Romas Kazlauskas and Claudia Schmidt-Dannert labs for generous use of their lab equipment.

Funding

We thank the U.S. Department of Energy Joint Genome Institute for synthetic DNA. The work conducted by the U.S. Department of Energy (DOE) Joint Genome Institute, a DOE

Office of Science User Facility, is supported under [DE-AC02-05CH11231]; The National Science Foundation Graduate Research Fellowship [00039202 to S.L.R.]; National Institutes of Health Biotechnology training grant [5T32GM008347-27 to M.D.S.]. We also acknowledge support from the MnDRIVE initiative for Industry and the Environment.

Conflict of interest statement. None declared.

References

- Nofiani, R., Philmus, B., Nindita, Y. and Mahmud, T. (2019) 3-Ketoacyl-ACP synthase (KAS) III homologues and their roles in natural product biosynthesis. *MedChemComm*, 10, 1517–1530.
- Haapalainen, A.M., Merilainen, G. and Wierenga, R.K. (2006) The thiolase superfamily: condensing enzymes with diverse reaction specificities. *Trends Biochem. Sci.*, 31, 64–71.
- Sukovich, D.J., Seffernick, J.L., Richman, J.E., Gralnick, J.A. and Wackett, L.P. (2010) Widespread head-to-head hydrocarbon biosynthesis in bacteria and role of OleA. *Appl. Environ. Microbiol.*, 76, 3850–3862.
- Christenson, J.K., Richman, J.E., Jensen, M.R., Neufeld, J.Y., Wilmot, C.M. and Wackett, L.P. (2017) β -Lactone synthetase found in the olefin biosynthesis pathway. *Biochemistry*, 56, 348–351.
- Frias, J.A., Richman, J.E., Erickson, J.S. and Wackett, L.P. (2011) Purification and characterization of OleA from *Xanthomonas campestris* and demonstration of a non-decarboxylative Claisen condensation reaction. *J. Biol. Chem.*, 286, 10930–10938.
- Bonk, B.M., Tarasova, Y., Hicks, M.A., Tidor, B. and Prather, K.L.J. (2018) Rational design of thiolase substrate specificity for metabolic engineering applications. *Biotechnol. Bioeng.*, 115, 2167–2182.
- Davies, C., Heath, R.J., White, S.W. and Rock, C.O. (2000) The 1.8 angstrom crystal structure and active-site architecture of beta-ketoacyl-acyl carrier protein synthase III (FabH) from *Escherichia coli*. *Structure*, 8, 185–195.
- Goblirsch, B.R., Jensen, M.R., Mohamed, F.A., Wackett, L.P. and Wilmot, C.M. (2016) Substrate trapping in crystals of the thiolase OleA identifies three channels that enable long chain olefin biosynthesis. *J. Biol. Chem.*, 291, 26698–26706.
- Goblirsch, B.R., Frias, J.A., Wackett, L.P. and Wilmot, C.M. (2012) Crystal structures of *Xanthomonas campestris* OleA reveal features that promote head-to-head condensation of two long-chain fatty acids. *Biochemistry*, 51, 4138–4146.
- Röttig, M., Rausch, C. and Kohlbacher, O. (2010) Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. *PLoS Comput. Biol.*, 6, e1000636.
- Chevette, M.G., Aicheler, F., Kohlbacher, O., Currie, C.R. and Medema, M.H. (2017) SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics*, 33, 3202–3210.
- Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C. and Kohlbacher, O. (2011) NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.*, 39, W362–W367.
- Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H. and Weber, T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, 47, W81–W87.
- Yang, M., Feh, C., Lees, K.V., Lim, E.K., Offen, W.A., Davies, G.J., Bowles, D.J., Davidson, M.G., Roberts, S.J. and Davis, B.G. (2018) Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat. Chem. Biol.*, 14, 1109–1117.
- Robinson, S.L., Terlouw, B.R., Smith, M.D., Pidot, S.J., Stinear, T.P., Medema, M.H. and Wackett, L.P. (2019) Global analysis of adenylate-forming enzymes reveals β -lactone biosynthesis pathway in pathogenic *Nocardia*. *bioRxiv*, doi: 10.1101/856955.
- Pethe, M.A., Rubenstein, A.B. and Khare, S.D. (2019) Data-driven supervised learning of a viral protease specificity landscape from deep sequencing and molecular simulations. *Proc. Natl. Acad. Sci. USA*, 116, 168–176.
- Chen, C.T., Yang, E.W., Hsu, H.J., Sun, Y.K., Hsu, W.L. and Yang, A.S. (2008) Protease substrate site predictors derived from machine learning on multilevel substrate phage display data. *Bioinformatics*, 24, 2691–2697.
- Song, J.N., Tan, H., Perry, A.J., Akutsu, T., Webb, G.I., Whisstock, J.C. and Pike, R.N. (2012) PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One*, 7, e50300.
- Engström, K., Nyhlén, J., Sandström, A.G., and Bäckvall, J.E., (2010) Directed evolution of an enantioselective lipase with broad substrate scope for hydrolysis of alpha-substituted esters. *J. Am. Chem. Soc.*, 132, 7038–7042.
- Smith, M.D., Robinson, S.L., Molomjams, M. and Wackett, L.P. (2020) *In vivo* assay reveals microbial OleA thiolases initiating hydrocarbon and β -lactone biosynthesis. *mBio*, 11, e00111–e00120.
- Cao, D.S., Xiao, N., Xu, Q.S. and Chen, A.F. (2015) Rcp: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, 31, 279–281.
- Cao, Y., Charisi, A., Cheng, L.C., Jiang, T. and Girke, T. (2008) ChemmineR: a compound mining framework for R. *Bioinformatics*, 24, 1733–1734.
- Wright, E.S. (2015) DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics*, 16, 322.
- Atchley, W.R., Zhao, J.P., Fernandes, A.D. and Drüke, T. (2005) Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA*, 102, 6395–6400.
- Osorio, D., Rondon-Villarreal, P. and Torres, R. (2015) Peptides: a package for data mining of antimicrobial peptides. *R J.*, 7, 4–14.
- Kuhn, M. (2008) Building predictive models in R using the caret package. *J. Stat. Softw.*, 28, 1–26.
- Wright, M.N. and Ziegler, A. (2017) Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.*, 77, 1–17.
- Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, e9490.
- Wallace, I.M. and Higgins, D.G. (2007) Supervised multivariate analysis of sequence groups to identify specificity determining residues. *BMC Bioinformatics*, 8, 135.
- Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.*, 10, 845–858.
- Tian, W., Chen, C., Lei, X., Zhao, J.L. and Liang, J. (2018) CASTp 3.0: computed atlas of surface topography of proteins. *Nucleic Acids Res.*, 46, W363–W367.
- Liang, G. and Li, Z. (2007) Factor analysis scale of generalized amino acid information as the source of a new set of

- descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides. *QSAR Comb. Sci.*, 26, 754–763.
33. Kidera, A., Konishi, Y., Oka, M., Ooi, T. and Scheraga, H. (1985) Statistical-analysis of the physical-properties of the 20 naturally-occurring amino-acids. *J. Protein Chem.*, 4, 23–55.
34. Sukovich, D.J., Seffernick, J.L., Richman, J.E., Hunt, K.A., Gralnick, J.A. and Wackett, L.P. (2010) Structure, function, and insights into the biosynthesis of a head-to-head hydrocarbon in *Shewanella oneidensis* strain MR-1. *Appl. Environ. Microbiol.*, 76, 3842–3849.
35. Allemann, M.N., Shulse, C.N. and Allen, E.E. (2019) Linkage of marine bacteria polyunsaturated fatty acid and long-chain hydrocarbon biosynthesis. *Front. Microbiol.*, 10, 702.
36. Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W. and Kopple, K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, 45, 2615–2623.
37. Huffman, M.A., Fryszkowska, A., Alvizo, O., Borra-Garske, M., Campos, K.R., Canada, K.A., Devine, P.N., Duan, D., Forstater, J.H., Grosser, S.T. et al., (2019) Design of an *in vitro* biocatalytic cascade for the manufacture of islatravir. *Science*, 366, 1255–1259.
38. Meng, G.Y., Guo, T.J., Ma, T.C., Zhang, J., Shen, Y.C., Sharpless, K.B. and Dong, J. (2019) Modular click chemistry libraries for functional screens using a diazotizing reagent. *Nature*, 574, 86–89.