MACHINE LEARNING AND BIOINFORMATIC INSIGHTS INTO KEY ENZYMES FOR A BIO-BASED CIRCULAR ECONOMY

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Engineering
at the University of Kentucky

By

Japheth E. Gado

Lexington, Kentucky

Co- Directors: Dr. Christina M. Payne, Adjunct Professor of Chemical and Materials Engineering

and Dr. Stephen Rankin, Professor of Chemical and Materials Engineering

Lexington, Kentucky

2020

https://orcid.org/0000-0002-0024-2531

ABSTRACT OF DISSERTATION

MACHINE LEARNING AND BIOINFORMATIC INSIGHTS INTO KEY ENZYMES FOR A BIO-BASED CIRCULAR ECONOMY

The world is presently faced with a sustainability crisis; it is becoming increasingly difficult to meet the energy and material needs of a growing global population without depleting and polluting our planet. Greenhouse gases released from the continuous combustion of fossil fuels engender accelerated climate change, and plastic waste accumulates in the environment. There is need for a circular economy, where energy and materials are renewably derived from waste items, rather than by consuming limited resources. Deconstruction of the recalcitrant linkages in natural and synthetic polymers is crucial for a circular economy, as deconstructed monomers can be used to manufacture new products. In Nature, organisms utilize enzymes for the efficient depolymerization and conversion of macromolecules. Consequently, by employing enzymes industrially, biotechnology holds great promise for energy- and cost-efficient conversion of materials for a circular economy. However, there is need for enhanced molecular-level understanding of enzymes to enable economically viable technologies that can be applied on a global scale. This work is a computational study of key enzymes that catalyze important reactions that can be utilized for a bio-based circular economy. Specifically, bioinformatics and data-mining approaches were employed to study family 7 glycoside hydrolases (GH7s), which are the principal enzymes in Nature for deconstructing cellulose to simple sugars; a cytochrome P450 enzyme (GcoA) that catalyzes the demethylation of lignin subunits; and MHETase, a tannase-family enzyme utilized by the bacterium, *Ideonella sakaiensis*, in the degradation and assimilation of polyethylene terephthalate (PET). Since enzyme function is fundamentally dependent on the primary amino-acid sequence, we hypothesize that machine-learning algorithms can be trained on an ensemble of functionally related enzymes to reveal functional patterns in the enzyme family, and to map the primary sequence to enzyme function such that functional properties can be predicted for a new enzyme sequence with significant accuracy. We find that supervised machine learning identifies important residues for processivity and accurately predicts functional subtypes and domain architectures in GH7s. Bioinformatic analyses revealed conserved active-site residues in GcoA and informed protein engineering that enabled expanded enzyme specificity and improved activity. Similarly, bioinformatic studies and phylogenetic analysis provided evolutionary context and identified crucial residues for MHET-hydrolase activity in a tannase-family enzyme (MHETase). Lastly, we developed machine-learning models to predict enzyme thermostability, allowing for high-throughput screening of enzymes that can catalyze reactions at elevated temperatures. Altogether, this work

provides a solid basis for a computational data-driven approach to understanding, identifying, and engineering enzymes for biotechnological applications towards a more sustainable world.

KEYWORDS: Machine learning, bioinformatics, enzymes, cellulase, protein thermostability, protein engineering

Japheth E. Gado

(Name of Student)

11/14/2020

Date

MACHINE LEARNING AND BIOINFORMATIC INSIGHTS INTO KEY ENZYMES FOR A BIO-BASED CIRCULAR ECONOMY

By
Japheth E. Gado

Dr. Christina M. Payne
Co-Director of Dissertation

Dr. Stephen Rankin
Co-Director of Dissertation

Dr. Stephen Rankin
Director of Graduate Studies

11/14/2020
Date

Dedication

To the vibrant community on stackoverflow.com, who provide the much-needed support that facilitates computational research around the world.

Acknowledgments

I would like to thank the funding source for my Ph.D. research, the National Science Foundation (CBET-1552355).

I am deeply grateful for my doctoral advisor, Dr. Christina M. Payne, for providing exceptional support and guidance throughout my Ph.D. program. By maintaining a consistently encouraging and assuring rapport, she always ensured that I was motivated towards excellence. I am thankful for the members of my dissertation committee: Dr. Steven Rankin, Dr. Tate Tsang, Dr. Brent Harrison, and Dr. Gregg Beckham for their useful advice on my research work. I am especially thankful to Dr Beckham, for accepting me as a collaborative researcher at the National Renewable Energy Laboratory (NREL), and ensuring that I had everything I needed for success. I would like to extend my appreciation to researchers NREL for the technical assistance they provided at several points in my research: Dr. Brandon Knott, Dr. Heather Mayes, Dr. Vivek Bharadwaj, Dr. Josh Vermaas, Dr. Peter St. John, and Mr. Benjamin Pollard. I am grateful for the opportunity to collaborate with Dr. Jerry Ståhlberg, Dr. Mats Sandgren, and Mr. Topi Haataja at the Swedish University of Agricultural Sciences, and I would like to specially thank them for their kind support.

Lastly, I appreciate my parents and sister for their love and encouragement throughout my program.

Table of Contents

List of Tables

List of Figures

.

CHAPTER 1. Introduction

1.1 Motivation

The global human population is continuously increasing, and is expected to surpass 11 billion by the year 2100.[1, 2] As steady population growth leads to a subsequent increase in demand for energy and materials, which will be accelerated by rapid urbanization in the developing world,[3] there is a critical need for technologies that meet the growing demand for resources without eroding Earth's life-support system. Regardless of future population growth and the increase in the demand for energy and materials, current technologies for meeting the needs of today's human population are markedly depleting limited natural resources and increasingly polluting the environment.[4-6]

Currently, more than 80% of the global energy demand is met by the combustion of fossil fuels, and carbon dioxide released from burning fossil fuels accounts for over 70% of total greenhouse gas emissions.[7] As the increase of greenhouse gases in the atmosphere is linked to global warming, drastically cutting back on fossil fuel consumption is of critical importance to limit anthropogenic climate change.[8] In addition to climate change, fossil fuels are limited and are not uniformly distributed among countries. As a result, energy dependence on fossil fuels is unsustainable and is associated with negative economic implications.[5] Moreover, petroleum is the leading source of majority of organic chemicals, which are utilized for their intrinsic value or as feedstock for producing diverse pharmaceuticals and synthetic materials. Dwindling petroleum reserves and pollution due to poor management of chemical waste are two key problems associated with dependence

on petroleum for chemicals and materials. In the United States alone, it is estimated that chemical industries release nearly 1.5 billion pounds of hazardous waste annually.[6]

The need for sustainable and environmentally friendly solutions to the growing demand for energy and materials necessitates a circular economy, in which energy and materials are continuously derived from renewable resources and waste items in a circular manner. In a circular economy, waste is virtually eliminated as industrial value is recouped from waste items by utilizing them as feedstock for the production of new resources. One of the most promising sustainable alternatives to petroleum is plant-derived material, commonly termed biomass. Biomass consists mostly of carbohydrates, such as starch and cellulose, and lignin, and these provide potential for the renewable production of nearly all primary chemicals derived from petroleum.[9] However, a major limitation to fully exploiting this potential is the marked resistance of biomass to chemical deconstruction.[10]

In Nature, microorganisms have evolved enzymatic machinery to efficiently deconstruct biomass. As such, enzymatic strategies for deconstructing biomass in an industrial context holds great promise for a circular economy. Moreover, many synthetic polymers, such as polyethylene terephthalate (PET) and polyurethane (PU), consist of similar chemical linkages as in several natural polymers that are efficiently deconstructed by microbial enzymes.[11] Thus, enzymes provide a cost- and energy-efficient way to deconstruct biopolymers and synthetic polymers into constituent monomers, which can be used to manufacture biofuels and new products. Consequently, gaining deeper mechanistic understanding of enzyme function will enable the engineering of high-yield enzymes and facilitate the journey towards a bio-based circular economy and a sustainable world.

The focus of this dissertation is to investigate relationships between the amino-acid sequence and catalytic function for key enzymes that are utilized in conversion of cellulose, lignin, and polyethylene terephthalate. Across the tree of life, evolutionary changes in protein sequence through mutation have given rise to a wide spectrum of homologous enzymes with varying functional properties. By a data-driven study of an ensemble of functionally related proteins, the deterministic relationships between amino-acid sequence and enzyme function can be elucidated, providing a more enhanced molecular-level knowledge of key enzymes that can be applied towards a bio-based circular economy. Herein, we apply bioinformatics and data mining techniques to large protein sequence datasets to map the variation in sequence to enzyme function, and to derive statistical relationships that can facilitate improving catalytic performance via protein engineering.

1.2 Research background

1.2.1 Cellulose and cellulases

Plant cell walls consist mostly of lignin, an aromatic polymer; and the polysaccharides, cellulose and hemicellulose, which constitute between 10-30%, 15-30%, and 20-50% of the dry weight plant cell walls, respectively.[12, 13] Cellulose is a homopolymer consisting of β-1,4-linked β-D-glucose units in a linear chain with a reducing and a non-reducing end. Due to the abundance of hydroxyl groups, cellulose forms numerous hydrogen bonds between oxygen atoms on the same chain or an adjoining chain (Figure 1.1). The network of hydrogen bonds, as well as van der Waals interactions, results in the organization of cellulose as crystalline microfibrils with great tensile strength. Unlike

cellulose, hemicellulose is a heteropolymer with random amorphous structure. The mechanical toughness of cellulose makes it an excellent biomaterial for conferring strength in plant cell walls. The rigidity of plant stems and tree wood results from the intricate organization of cellulose fibers with lignin to form a lignocellulosic complex.[14] In addition to plant cell walls, cellulose is utilized as a structural component in many algae, oomycetes, and bacteria. Consequently, cellulose is the most abundant biopolymer on earth.[15]

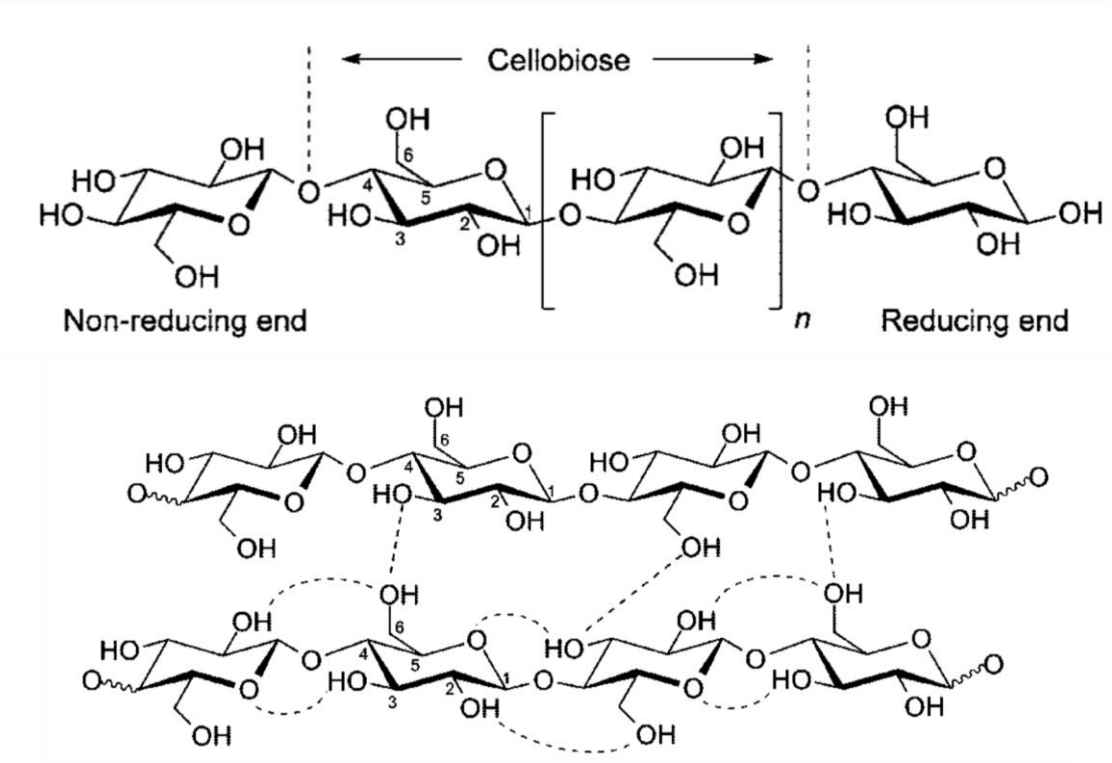


Figure 1.1 Chemical structure of cellulose, showing the organization of glucose units into a linear chain with a reducing and non-reducing end (above), and hydrogen bond linkages between two cellulose chains. This figure has been adapted with permission from Pinkert et al.,[16] copyright 2009, American Chemical Society.

The crystalline structure of cellulose may exist in different forms, known as polymorphs. Cellulose produced by natural systems, such as plants, bacteria, and algae, are of type cellulose I, which consists of parallel-oriented chains with intralayer and interlayer hydrogen bonds but no intersheet hydrogen bonds.[17] Hence, van der Waals forces between sheets are inferred to contribute significantly to the overall stability of cellulose I structures.[12] There are two types of cellulose I: cellulose Iα and Iβ. The major difference between Iα and Iβ is the pattern of intralayer hydrogen bonding and the nature of interlayer chain stacking. Whereas Iα packs to form a unit cell with a single chain, Iβ forms two layers (Figure 1.2).[18, 19] Plant cellulose has been shown to be a mixture of both Iα and Iβ.[20, 21] Cellulose II and III are derived from chemical treatment of cellulose I, and are, thus, synthetic polymorphs. Unlike cellulose I, cellulose II and III form antiparallel or staggered layers with interlayer hydrogen bonds, and are more susceptible to enzymatic degradation.[22, 23]

Cellulose is a remarkably recalcitrant polymer. On the molecular level, strong covalent glycosidic linkages, hydrogen bonds between chains, and hydrophobic interactions between sheets make cellulose profoundly resistant to chemical hydrolysis. Furthermore, the unique organization of the lignocellulosic complex in plant tissues limits liquid penetration and enzyme accessibility.[10] However, across the tree of life, organisms have evolved mechanisms for the reorganization and conversion of cellulose to soluble constituents as energy source. Numerous cellulolytic microbes produce enzymes that act synergistically to deconstruct lignocellulosic material in plant cell wall. Broadly, three types of enzymes are employed to hydrolyze cell wall matter: cellulases, hemicellulases, and accessory enzymes.[10, 12] Upon overcoming the hemicellulose barrier in cell-wall

microfibrils, cellulases and accessory enzymes are employed to deconstruct the cellulose core of cell-wall matter.

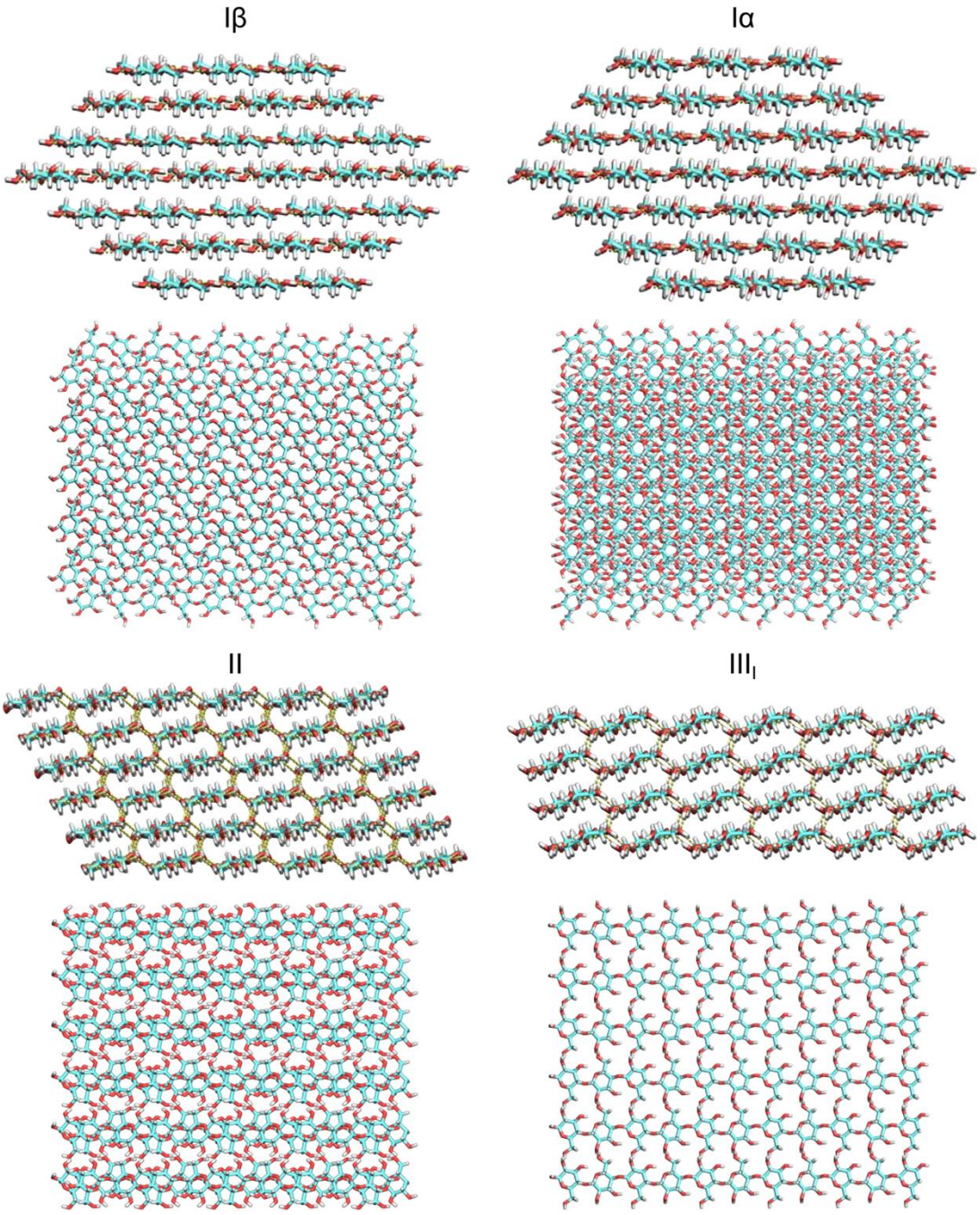


Figure 1.2 Structure of cellulose polymorphs. Naturally occurring polymorphs (Iα and Iβ) exhibit only interlayer hydrogen bonding. Synthetic polymorphs, II and III_I, which are

derived from chemically treatment of cellulose I, exhibit hydrogen bonding between layers. Hydrogen bonding is shown in yellow. This figure was reprinted with permission from Payne et al.,[12] copyright 2015, American Chemical Society.

Microbial cellulose deconstruction is achieved via free enzymes or cellulosomes, in which large complexes of hundreds of enzymes held by noncovalent interactions operate in close proximity.[24] Free cellulases include cellobiohydrolases, endoglucanases, and β-glucosidases. Cellobiohydrolases attach to free cellulose chain ends and processively cleave multiple cellobiose units as they thread along the chain from the reducing end towards the non-reducing end, or vice versa. Cellobiohydrolases were once thought to be purely exo-acting and were previously called exoglucanases, but studies have shown that they are capable of endo-initiation as well.[25, 26] Endoglucanases attack internal bonds in amorphous regions of cellulose, which is helpful to make open chain ends accessible to cellobiohydrolases. β-glucosidases are accessory enzymes that hydrolyze the cellobiose products of cellobiohydrolases to single glucose units and facilitate the reaction process by reducing product-inhibition. The prevailing paradigm of cellulose degradation, in Nature and industrially, involves the use of a synergistic cocktail of cellobiohydrolases, endoglucanases, and accessory enzymes such as β-glucosidases and lytic polysaccharide monooxygenases (LPMOs). LPMOs utilize a metal-dependent oxidative mechanism to cleave glycosidic bonds and are capable of creating chain breaks in difficult crystalline regions of cellulose that are inaccessible to endoglucanases.[27]

Among cellulolytic organisms, fungi are responsible for the bulk of cellulose degradation in nature, and many filamentous fungi are capable of secreting markedly large amounts of cellulolytic free enzymes. Cellobiohydrolases, being capable of successively

cleaving multiple cellobiose units before dissociating form the substrate, are responsible for the majority of hydrolytic bond cleavages. Consequently, fungal cellobiohydrolases are the most promising targets for scientific studies and protein engineering towards developing more efficient and cost-effective industrial technologies for biomass conversion.

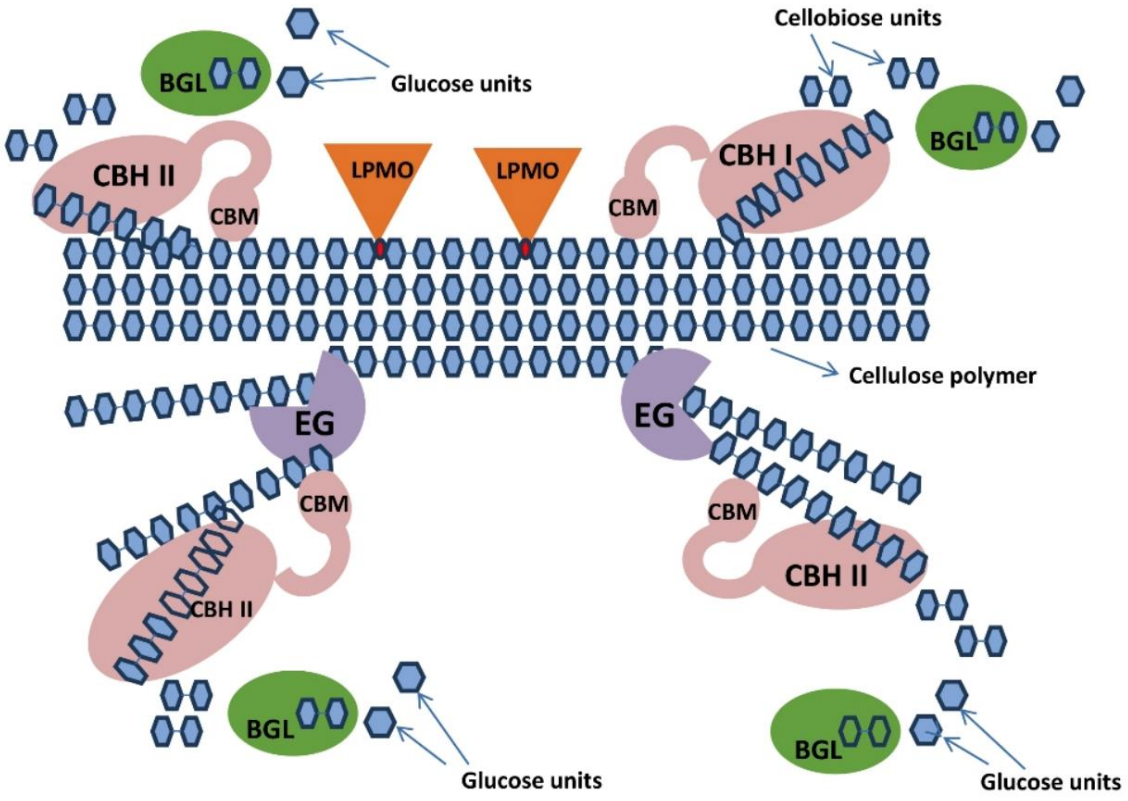


Figure 1.3 Synergistic action of free enzymes for deconstruction of cellulose polymer. Endoglucanases (EG) hydrolyze amorphous regions within the chain and lytic polysaccharide monooxygenases (LPMO) oxidize crystalline regions within the chain. Cellobiohydrolases (CBH I and CBH II) processively cleave off cellobiose units as they attach to open chain ends made available by EGs and LPMOs and thread through the

cellulose chain. β-glucosidases (BGL) hydrolyze cellobiose to glucose units. This figure was reprinted with permission from Singh et al,[28] copyright 2017, Biofuel Research Journal.

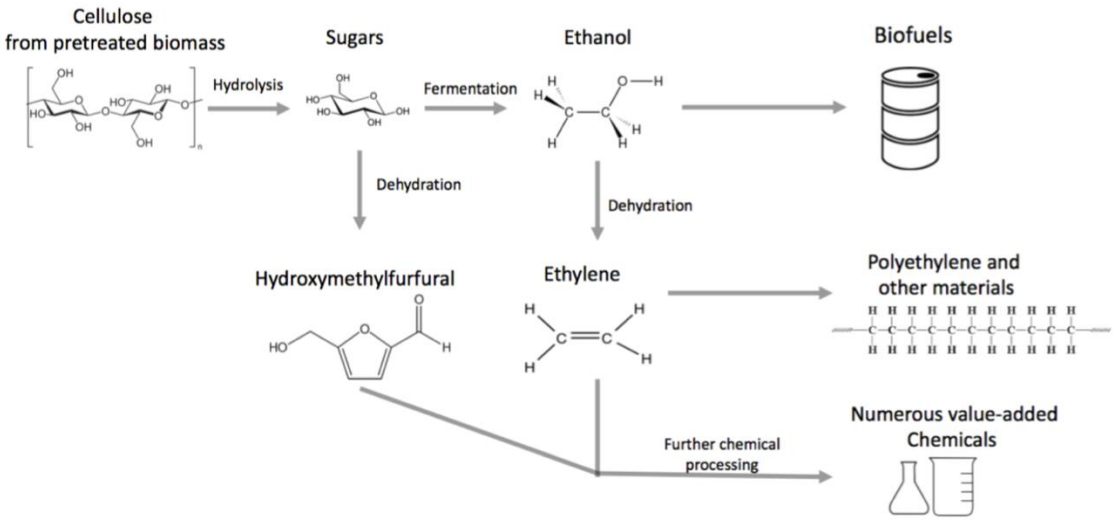


Figure 1.4 A schematic description of conversion of cellulose to high-value products.

With natural global production of cellulose exceeding 180 billion tons annually,[29] employing high-yield cellulases industrially provides a sustainable means to tap into the massive wealth available in lignocellulosic biomass and markedly offset dependence on petroleum. Sugars derived from cellulose deconstruction can be converted to ethanolic biofuels and a wide variety of high-value chemicals through intermediate platforms such as ethylene and hydroxymethylfurfural (Figure1.4).[30]

1.2.2 Family 7 glycoside hydrolases

Glycoside hydrolases (GH) are enzymes that catalyze the hydrolysis of glycosidic bonds and include enzymes such as cellulases, chitinases, amylases, galactosidases, mannosidases, etc. The Carbohydrate-Active enZYmes database (CAZY) classifies GHs into families according to amino-acid sequence similarities.[31] Currently, there are 168 families and several catalytic activities may be present in each family. For example, family 1 glycoside hydrolases (GH1) demonstrate β-glucosidase (EC 3.2.1.21), β-D-fucosidase (3.2.1.38), lactase (EC 3.2.1.108), vicianin hydrolase (EC 3.2.1.119), and several other enzyme activities. Furthermore, the same enzyme activity may appear in several families. Endoglucanase activity (EC 3.2.1.4) is found in more than 15 families.

Glycoside hydrolysis mechanism can be generally described as either retaining or inverting.[32] While inverting mechanism is a one-step reaction, retaining mechanism is a two-step reaction. The single catalytic step in inverting mechanism involves a nucleophilic attack at the anomeric carbon of the saccharide by a water molecule in which a proton is transferred from the water molecule to the catalytic base and the glycosidic bond is broken due to the transfer of a proton from the catalytic acid.[12] The first step in retaining mechanism, the glycosylation step, is an attack at the anomeric carbon by the nucleophile following a proton transfer from the catalytic acid. This step results in the formation of a glycosyl-enzyme intermediate and the inversion in the stereochemistry of the sugar. The second step is the deglycosylation step in which the anomeric carbon is attacked by a nucleophilic water molecule breaking the glycosyl-enzyme intermediate bond followed by a transfer of a proton to the catalytic base which inverts the stereochemistry back to its original state.[12] GHs typically follow either one of retaining or inverting mechanisms,

although several GH families have been shown to deviate from this paradigm.[32, 33] The catalytic acid, the catalytic base (in inverting mechanism), and the nucleophilic residue (in retaining mechanism) have carboxylate groups.

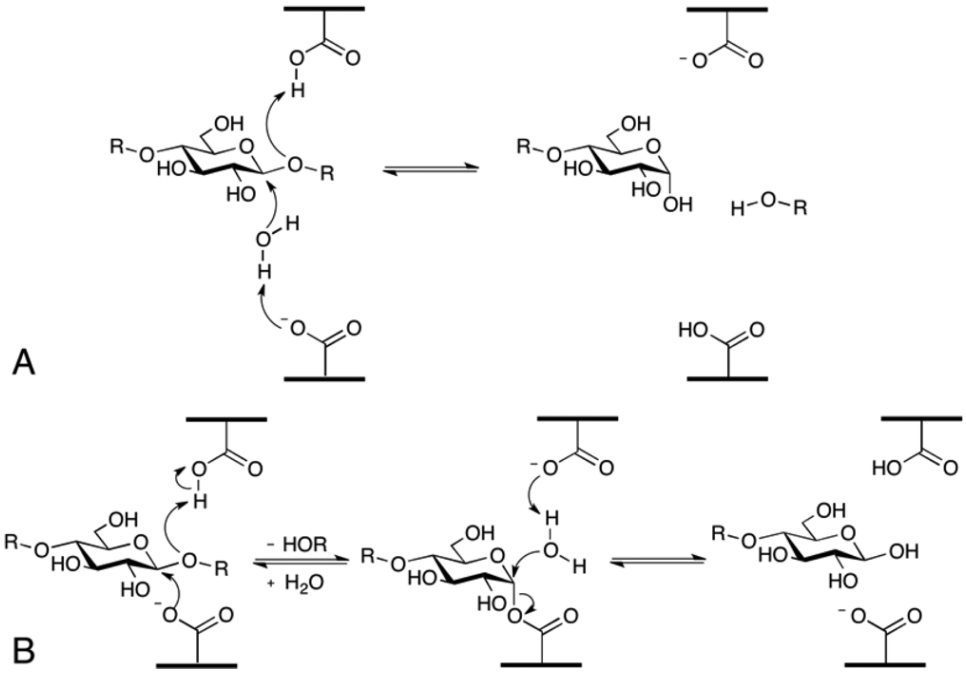


Figure 1.5 Two possible mechanisms for glycosidic hydrolysis. (A) Inverting mechanism showing the nucleophilic attack by a water molecule on the anomeric carbon and the transfer of a proton from the acid to the glycosidic oxygen. (B) Retaining mechanism occurring in two steps: the glycosylation step, in which a glycosyl-enzyme intermediate is formed by a proton transfer from the acid to the glycosidic oxygen, and the deglycosylation step, in which water attacks the anomeric carbon, a proton is transferred to base, and the enzyme is restored to its initial state. This figure was reprinted with permission from Payne et al.,[12] copyright 2015, American Chemical Society.

Among GHs that cleave β-1,4-glycosidic bonds in cellulose (cellulases), family 7 glycoside hydrolases (GH7) are the chief enzymes for cellulose degradation, both in nature and in industrial processes. All known cellulolytic fungi utilize GH7s, and GH7s are often the predominant enzymes by mass in their secretomes (up to 50%).[34, 35] As fungi are the powerhouses of cellulose degradation in nature, GH7s play a critical role in the carbon cycle. Although GH7s are found majorly in fungi, they have been identified in several non-fungal species like Crustacea, Porifera, Parabasalia, Alveolata, and Amoebozoa. To date, GH7s have not been found in bacteria.

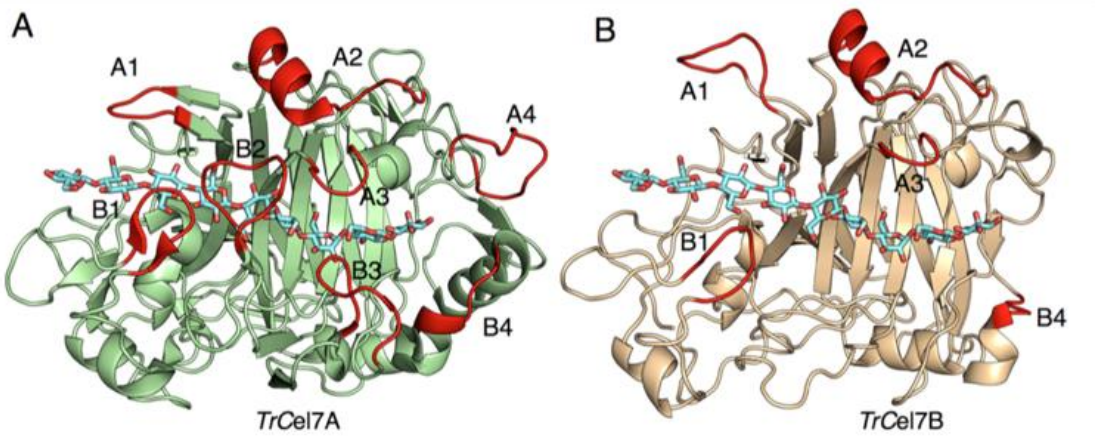


Figure 1.6 Example structures of a (A) GH7 CBH (*Tre*Cel7A) and (B) a GH7 EG (*Tre*Cel7B) from *Trichoderma reesei* with a 9-member cellulose chain in the active site. The eight active-site loops are named A1 to A4 and B1 to B4, some of which are markedly truncated in the EG, resulting in a more open active site. This figure was reprinted with permission from Payne et al.,[12] copyright 2015, American Chemical Society.

GH7 enzymes are either CBHs or EGs and adopt a retaining mechanism with a strongly conserved catalytic triad of Glu and Asp residues in an EXDXXE motif. GH7

CBHs and EGs share a similar β-jelly roll fold with a two antiparallel β-sheets packing into a curved β-sandwich. The major differences in their structures lies in the active-site region, which is more open or cleft-like in GH7 EGs due to truncation in loops that protrude over the active site, and more closed or tunnel-like in GH7 CBHs due to longer active-site loops (Figure 1.6). GH7 CBHs processively cleave cellulose chains from the reducing end towards the non-reducing end of the chain while GH7 EGs are endo-acting and are mostly non-processive. The processivity of GH7 CBHs, in addition to being secreted by cellulolytic fungi in large amounts, make them a focus of scientific studies and in industrial applications.[36] However, there are still gaps in understanding how GH7 sequence relate to functional attributes, such as processivity and activity. About a third of GH7 enzymes are attached to a carbohydrate binding module (CBM) as an auxiliary domain. It is now generally accepted that the CBM does not directly affect hydrolysis directly, but functions majorly to improve the binding affinity of the catalytic domain for the cellulose substrate.[37, 38]

Despite the wealth of diverse sequence and structural data available, most studies of GH7s have involved experimentation and comparative analysis of only a few GH7s.[39-44] As a result, the wealth of information available through studies of a large ensemble of proteins—which uncover patterns in the evolutionary design of GH7 diversity across the eukaryotic tree of life—remains untapped. The need for improved cellulases with lower product inhibition,[45] higher catalytic efficiency,[46] and elevated thermostability,[47, 48] for accelerated industrial processes, call for integrating a data-driven, family-wide approach to studying GH7s, in addition to traditional rational-design approaches.

1.2.3 Enzymatic demethylation of lignin subunits

Lignin is a complex heteropolymer of aromatic subunits. In many plants, the cell walls contain a matrix formed from polysaccharides and lignin that provides enhanced rigidity and strength. Additionally, the abundance of aromatic groups in lignin enhances the hydrophobic nature of the cell wall, making the cell water impermeable and resistant to microbial and chemical attack. Trees utilize vast amounts of lignin for structural support, with up to 36% of the dry weight of wood comprising of lignin.[49] As a result, lignin is one of the world's most abundant natural polymers and provides the largest renewable source of aromatic carbon in nature. Despite the abundance of lignin in the environment and its potential to replace non-renewable petroleum as the main source of aromatic chemicals and materials in a bio-based economy, lignin has received little industrial attention for deriving economic value and valorization.[50, 51] In paper production, lignin is removed from lignocellulose before papermaking. Although over 50 million tons of lignin is extracted yearly by the paper industry, less than 2% of it is used commercially as low-value chemicals and the rest is burned as low-value fuel.[50, 52] There is therefore need to develop industry-scale technology for the conversion of lignin to high-value products.

Despite the recalcitrance of lignin, many fungi and bacteria have evolved powerful enzymatic systems that deconstruct lignin into smaller chemical fragments.[53-55] Lignin is broken down in Nature by enzymes, such as peroxidases, laccases and additional oxidative enzymes, that produce aromatic radicals which attack the various linkages in lignin via non-enzymatic reactions.[51] These reactions result in a wide variety of smaller aromatic fragments which are further cleaved and assimilated by the microorganisms as carbon and energy source via several aromatic-catabolic pathways.[56, 57] The enzymatic strategies

employed by microbes for the deconstruction and assimilation of lignin offer a potential for the valorization of lignin to high-value chemicals.[51]

Lignin is primarily composed of three aromatic monomeric units (or monolignols) that differ in the substitution patterns of the aromatic ring and the degree of methoxylation: p-coumaryl alcohol (H), coniferyl alcohol (G), and sinapyl alcohol (S).[58] The relative abundance of these monolignols varies from species to species. In some species, additional subunits form a relatively significant amount of monolignols, including hydroxycinnamic acids, caffeoyl alcohol, tricine, resveratrol, isorhapontigenin, and hydroxystilbene glucosides.[59-63] In the microbial deconstruction and metabolism of lignin, a crucial chemical step is the demethylation of lignin-derived compounds to diols before they are cleaved to ring-opened compounds.

Due to the wide diversity of methoxylated lignin products, there is need for discovery of demethylase enzymes in nature that can be adapted and engineered for industrial applications. In recent years, a few microbial enzymes have been discovered and studied that are capable of demethylating the methoxy group in a number of lignin substrates such as vanillate, 3-O-methylgallate, syringate, guaiacol, and guaethol.[64-69] A notable development was the discovery of the cytochrome P450-reductase gene pair (*gcoAB*) from *Amycolatopsis sp*. ATCC 39116 that showed demethylase activity on several lignin-derived products, including guaiacol, guaethol, 3-methyl-catechol, anisole, and 2-methyl anisole.[70] The promiscuity of the GcoAB enzyme system provides a particular advantage for biotechnological applications, as the same enzyme system can be used to catalyze the demethylation of a heterogenous stream of lignin-derived products from an upstream process. However, GcoAB did not show detectable activity on vanillate, ferulate

and veratrole. The promiscuous demethylase activity of GcoAB on several substrates suggests a notably flexible active site that can be expanded to accommodate other lignin-substrate specificities via protein engineering. A data-driven approach that examines the amino-acid variation at active-site positions across the family of homologs (family CYP 255A) is a promising way to discover key positions that can be mutated to expand the substrate specificity of GcoAB.

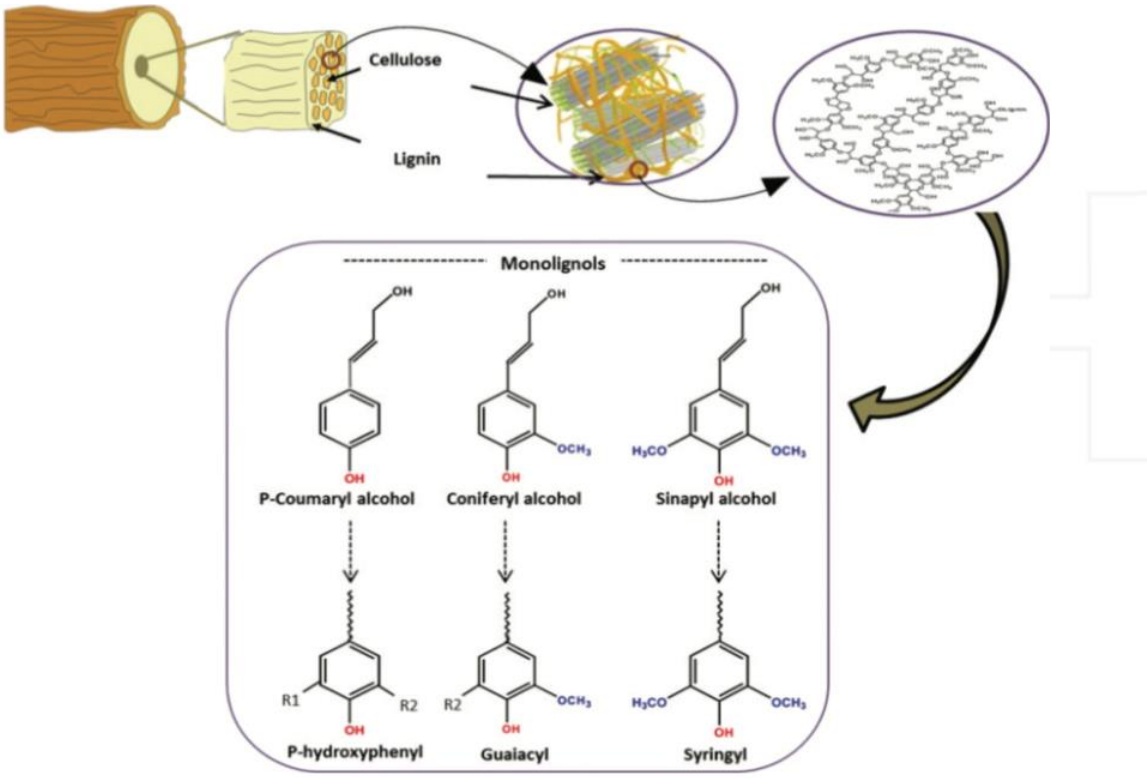


Figure 1.7 Lignin as a major component of biomass, existing in a matrix with cellulose. The three major monomeric components of lignin (monolignols) and the respective lignin products derived from them are also shown. This figure was reprinted from Mandlekar et al,[71] copyright 2018.

1.2.4 Enzymatic degradation of polyethylene terephthalate

Polyethylene terephthalate (PET) is a synthetic polymer derived from terephthalic acid and ethylene glycol. PET is a polyester desired for its unique properties, such as light weight, mechanical strength, durability, cheapness, and non-degradability. As a result, PET is the most common thermoplastic polyester plastic on earth today, being widely used in packaging food and drinks and in fabric fibers. It is estimated that about 56 million tons of PET is produced annually worldwide.[72] Due to the large aromatic composition of PET (from the terephthalate group), PET is notably chemically inert and resistant to chemical and microbial degradation.[73] Efforts at managing PET waste are focused on mechanical recycling processes. However, the contamination of plastic waste streams and the huge energy requirements severely limit the successful application of mechanical recycling on the bulk of plastic waste generated. Due to the systematic difficulties associated with recycling, less than 10% of all plastics produced are recycled and nearly 80% ends up accumulating in landfills or the natural environment.[11, 74] Moreover, recycling results in lower quality products that ultimately end up being disposed or incenerated.[75, 76] Hence, there is a critical need for sustainable strategies that restart the plastic lifecycle and recoup value from plastic waste in a circular economy.

Similar to industry-scale microbial degradation of lignocellulose, researchers have long been interested in enzymatic strategies for deconstructing plastic polymers. Since synthetic polymers have similar chemical bonds as natural polymers, employing microbial enzymes in industrial processes to deconstruct plastic waste is a promising approach. Indeed, several microbial enzymes have been discovered that are capable of cleaving the ester bond in PET.[72, 77-79] These enzymes are typically cutinases or lipases from the ab-

hydrolase superfamily. Yoshida et al isolated a bacterium (*Ideonella sakaiensis* 201-F6) from soil samples around a PET recycling factory and observed that the bacterium is able to utilize PET as its major energy and carbon source and achieve complete degradation of amorphous PET.[72] *Ideonella sakaiensis* degrades PET by secreting two main enzymes. The first enzyme is a cutinase-like enzyme, named PETase, which attacks PET to produce bis(2-hydroxylethyl) terephthalate (BHET), mono(2-hydroxylethyl) terephthalate (MHET) and terephthalic acid (TPA). PETase also cleaves BHET to MHET but shows virtually no activity on MHET. The second enzyme, a tannase-family protein called MHETase, hydrolyzes MHET to produce TPA and ethylene glycol (EG). Thus, by a synergistic action of PETase and MHETase, *Ideonella sakaiensis* deconstructs PET to TPA and EG. Subsequently, TPA and EG are assimilated via catabolic pathways. TPA is transported by TPA transporter protein (TPATP) and catabolized to protacatechuic acid (PCA) by TPA 12,-dioxygenase (TPADO), both of which are notably upregulated when *Ideonella sakaiensis* was cultured on TPA-Na, BHET, or PET films.[72] As TPA and EG are derived from petroleum in the chemical industry, PETase and MHETase provide a promising biotechnological approach to offset petroleum dependence and manage plastic pollution by the efficient deconstruction of PET waste to the chemical building blocks (TPA and EG), which can be used to produce new PET material without a compromise of mechanical quality (as is experienced in mechanical recycling).

Several studies have focused on structural and biochemical characterization of PETase and other PET-hydrolase homologs, and on protein engineering to expand the specificity, improve catalytic efficiency, or enhance thermal stability.[80-84] However, fewer studies have been done on MHETase.[85, 86] As PETase does not hydrolyze MHET, it is likely

that PETase acts synergistically with MHETase to convert PET to TPA and EG. Thus, there is need to gain mechanistic understanding of MHETase and how MHET-hydrolase activity evolved in the tannase family of enzymes to facilitate engineering higher PET-degradation yields. For much greater yields, thermostable PETase variants that are capable of PET breakdown at elevated temperatures near the glass transition temperature of PET are desired. Machine learning provides an effective way for the high-throughput screening of PETase homologs to identify prospective thermostable PETases.

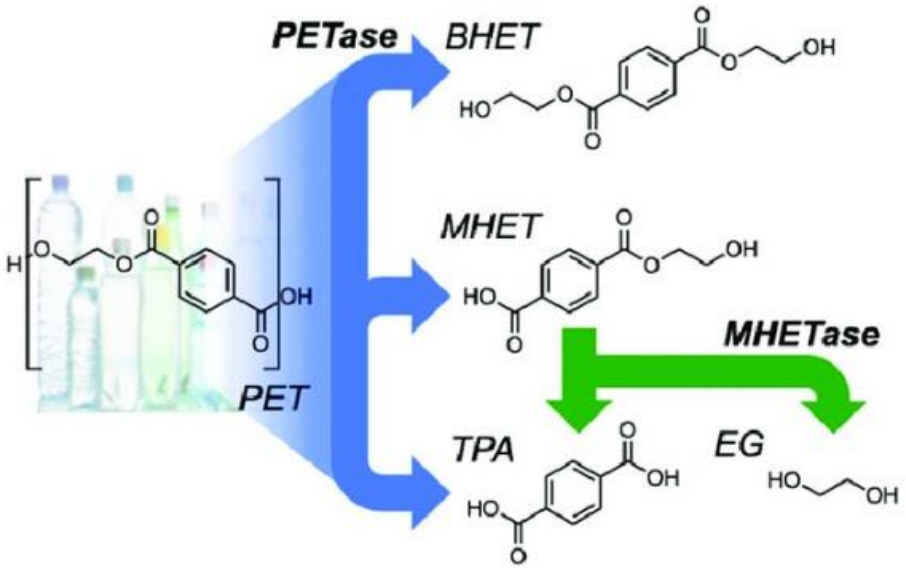


Figure 1.8 Combined action of the enzymes, PETase and MHETase, in *Ideonella sakaiensis* 201-F6 to deconstruct PET into TPA and EG. PETase adheres to PET and cleaves it to liberate BHET, MHET and TPA. PETase also cleaves BHET to yield MHET. MHETase hydrolyzes MHET to TPA and EG. This figure was reprinted with permission from Austin et al,[81] copyright 2018, National Academy of Sciences.

1.3 Outline of dissertation

The overall theme of this dissertation is to understand develop a fundamental understanding from a data-driven perspective of the relationships between the amino-acid sequence and functional variation of enzymes that catalyze the degradation of cellulose (family 7 glycoside hydrolases), lignin products (GcoA), and polyethylene terephthalate (PET). We hypothesize that statistical tools, such as machine learning, can be applied to an ensemble of homologous proteins to discover unique sequence and structural trends in the enzyme family of that can be exploited to manipulate the catalytic activity and thermostability.

The first part of this dissertation (Chapters 3 and 4) focuses on enzymes that are applicable for the degradation lignocellulosic biomass. In Chapter 3, machine learning is applied to discriminate between GH7 functional subtypes (i.e., CBH and EG) and to identify residue positions features that strongly correlate with subtype and that are, consequently, promising engineering targets. Machine learning is also implemented to predict the presence of a CBM domain in GH7s from residues in the catalytic domain. Key residues in the CD domain that correlate with the presence of a CBM were identified by determining positions that yielded highest feature importance in the random forest model. In Chapter 4, bioinformatic (conservation) analysis was applied to an ensemble of GcoA homologs to investigate the variability of active-site residues across the cytochrome P450 255A (CYP255A) family. The bioinformatic insights derived from the conservation analysis provided a basis for protein engineering efforts (by collaborators) to expand the substrate specificity and catalytic activity of the promiscuous GcoA enzyme to other lignin products.

The second part of this dissertation (Chapter 5) describes a large collaborative work on the characterization and engineering of a novel two-enzyme system for PET depolymerization. Bioinformatic and phylogenetic analyses of selected MHETase homologs were conducted to gain insight into the evolution of MHET-hydrolase activity in the tannase family. Conservation analysis of the residues within the coordination sphere of the MHET-substrate in the active site highlighted key residues in MHETase that notably differ from other tannase family sequences, and that play major roles in MHET-hydrolase activity. Biochemical studies, protein engineering, and molecular dynamics simulations (conducted by collaborators) provided insight into the reaction mechanism of MHETase and confirmed the functional roles of key residues identified by conservation analysis in MHETase and close homologs.

The final part of this dissertation (Chapters 6 and 7) focuses on machine learning for predicting protein thermostability. The goal was to develop machine learning models that can identify thermophilic enzymes that are active at high temperatures. We hypothesized that, since protein folding and structure is a deterministic function of the amino-acid sequence, machine learning models can be trained on the protein sequence alone to predict protein thermostability with significant accuracy. In Chapter 6, we present a support vector machine (ThermoProt) trained on a set of 32,000 diverse proteins to discriminate between psychrophilic, mesophilic, thermophilic, and hyperthermophilic proteins. In Chapter 7, we markedly improve on a machine learning method (TOME) for directly predicting the enzyme catalytic optimum temperature directly from the amino-acid sequence. By incorporating resampling strategies and ensemble learning to mitigate the

effects of data imbalance due to the skewed distribution of the training data, our new method (TOMER) achieves superior performance, particularly on high-temperature values.

CHAPTER 2. Computational Methodology

2.1 Introduction

We implemented supervised machine learning to predict enzyme functional attributes including activity subtype, domain architecture (i.e., the presence of a carbohydrate binding module attached to the catalytic domain), and optimal catalytic temperature. Analysis of the machine learning results further revealed key residues in the enzyme family that appear be related to function. Conservation analysis was implemented to investigate amino-acid variability in active-site positions and to gain insights for expanding substrate specificity through protein engineering. Phylogenetic analysis was performed to understand the evolution of MHET-hydrolase activity in the tannase family and to identify key positions in MHETase that relate the MHET substrate specificity and activity relative to other tannase-family sequences. Descriptions of the methods used are provided below.

2.2 Machine learning

Machine learning is the study and application of computer algorithms that are capable of learning to optimize their performance on a task from data or past experience.[87, 88] The unique characteristic of machine learning algorithms is that they are not explicitly programmed to solve a particular task. Rather than hard coding specific rules to solve a problem, machine learning algorithms solve problems by learning directly from the data. Machine learning algorithms discover existing relationships in the data by tuning a set of parameters to optimize a defined criterion. The central hypothesis of machine learning is

that if there is sufficient data, computer algorithms can learn the structure in the data, which can be exploited to solve a wide variety of problems. For example, in spam filtering, one could explicitly program an algorithm to identify spam emails by identifying key words in spam emails. But this approach would be rather cumbersome and inefficient. However, with sufficient examples of legitimate and spam emails, a machine learning algorithm can be trained on the data to implicitly learn the patterns that constitute a spam mail.[89] Although what constitutes spam mail may change with time, a machine learning algorithm can adapt to these changes when retrained on newly available data, without having to modify the algorithm.

Over the last few decades, machine learning has gained accelerated popularity in many fields, and has been employed to provide cutting-edge solutions to a wide variety of problems. Using machine learning, numerous artificial intelligence (AI) technologies that enable computers to make intelligent human-like decisions or recommendations have emerged. Notable examples include image recognition,[90] machine language translation,[91] medical diagnosis,[92] traffic prediction,[93] targeted advertising (i.e., predict which customers are most likely to respond to an ad),[94] and pattern recognition in scientific research.[95] In biology, particular in molecular biology, with the increasing growth of sequence and structure data, the practicality of machine learning is evident, and is demonstrated by numerous studies.[96-101] Machine learning algorithms can be applied to learn the complex, non-linear relationships between the functional attributes of biomolecules (DNA, proteins, etc.) and their sequence or structural components, allowing for predictive models of function or to gain deeper mechanistic knowledge. Deep learning is a special group of machine learning methods based on neural networks (connected computational graphs that

mimic neurons in the human brain), with many layers in the network. Compared to other traditional machine learning methods, deep learning allows representations to be learned directly from the data without the need for feature engineering or feature selection, and is capable of learning much deeper, less biased representations of the data, providing significant improvement in performance.[102] Deep learning has been the basis of the most recent breakthroughs in artificial intelligence.[103] However, the scarcity of labeled data in bioinformatics often limits the improvement in performance of deep learning models over traditional machine learning models.[102, 104, 105]

2.2.1 Common terminology in machine learning

A *feature* is an individual, measurable property of a data sample being analyzed that describes the sample and serves as input to the machine learning algorithm. A *target* or *label* is the final output variable of the sample that is being predicted. In *supervised learning*, data samples have both features and labels, and the goal is to learn a function that maps the features to the target/label such that the unknown target of a sample can be predicted from its known features. In *unsupervised learning*, however, data samples are not labeled, and the goal is to deconstruct hidden patterns and algorithmic relationships that are inherent in the data. If the targets are discrete classes, the supervised learning task is a *classification* problem. Otherwise, if the targets are continuous variables, the task is a *regression* problem.

Let X_m represent the m^{th} sample in a dataset described by n features such that $X_m = [x_m^1, x_m^2, x_m^3, \ldots, x_m^n]$, and let y_m represent the target label of the sample. The goal of a supervised machine learning algorithm, in mathematical terms, is to learn a function, H,

that maps the features of the data, *X*, to the labels, *y*, as accurately as possible. Generally, the process of learning *H* involves tuning a set of configuration values of the model, called *parameters*, such that the model function minimizes the error, ε.

$$y = H(X) + \varepsilon = \hat{y} + \varepsilon \tag{2.1}$$

A model may overfit the data it is trained on so that although it shows high performance in training, when applied to new data that were not seen in training, it fails to generalize. Such a model is said to have *high variance*, as it learns noise in the data and is overly sensitive to any perturbances in the data. On the contrary, a model may not be complex enough to learn patterns in the data so that it performs poorly both in training and on new data. Such a model is said to have *high bias*. The data used in machine learning are customarily divided into two or more sets. The *training set* is used to tune the parameters of the model in the learning process, and the *testing set* is used to derive an unbiased estimate of the model's performance. There may be certain values or options that the user must specify which are not learned be model from the data. These values are called *hyperparameters*. It is bad practice to use the testing set to optimize the choice of hyperparameters, as this may lead to overfitting the hyperparameters to the testing set and a consequent overestimation of the model performance. In such a case, it is advisable to split the data into three sets, including a validation set, which is used to select optimal hyperparameters so that the testing set is not touched until a final estimate is to be obtained.[106]

A more common approach is to first split the data into a training and testing, leave the testing set untouched until the end of the learning process, split the training set into several (k) folds, and repeatedly use one fold for validation and $k - 1$ folds for training until

all folds have been used in training and validation. The validation performance is determined as an average over all *k*-folds. This technique is called *k-fold cross-validation*. To minimize the variance due to the randomness of splitting the training set into folds, a random split may be performed *r* times, and the validation performance is determined as an average over all $r \times k$ folds. This technique is called *repeated k-fold cross-validation*.[107]

2.2.2 Feature scaling

Machine learning algorithms may be adversely affected by varying magnitudes of the features and target labels. Hence, it is often essential to perform *feature scaling* to monotonically transform each feature in the data so that they have a similar range or distribution. This is particularly important when using algorithms that depend on distances between the features. Widely varying magnitudes may cause some features to have a greater undesired impact on the outcome than others. Furthermore, several machine algorithms converge to a solution much faster with scaled features than with unscaled features. The following are popular feature scaling techniques:[108]

2.2.2.1 Min-max scaling (normalization)

This scales the feature to a desired range bound between minimum and maximum values, (*a* and *b*, respectively). Common boundary ranges are [0,1] and [-1,1]. Min-max scaling is sensitive to outliers in the data. In equation 2.2 below, x^i_{max} and x^i_{min} are the maximum and minimum values of feature *i* in the dataset.

$$x^i_{new} = \frac{x^i_m - x^i_{min}}{x^i_{max} - x^i_{min}} \times (b - a) + a \qquad (2.2)$$

2.2.2.2 Standard scaling (standardization)

This centers the feature distribution around a mean of 0 and standard deviation of 1. Given the mean and standard deviation of feature i, respectively, standard scaling is performed according to equation 2.3.

$$x_{new}^{i} = \frac{x_{m}^{i} - \mu^{i}}{\sigma^{i}} \tag{2.3}$$

2.2.2.3 Unit vector scaling

This scales the feature by dividing each value by the L2 norm so that the whole feature vector is of unit length.

$$x_{new}^{i} = \frac{x_{m}^{i}}{\|x^{i}\|} \tag{2.4}$$

2.2.2.4 Robust scaling

The disadvantage of all three above-mentioned scaling techniques is that these scaling methods are consequently significantly affected by severe outliers, since the range, mean, standard deviation, and L2 norm are skewed by outliers. A more robust method is to scale the feature to the interquartile range, which is less impacted by outliers. Equation 2.5 describes the scaling of the i^{th} feature to the range, [a,b] using the 1^{st} quartile (Q1) and third quartile (Q3).

$$x_{new}^{i} = \frac{x_{m}^{i} - Q1(x^{i})}{Q3(x^{i}) - Q1(x^{i})} \times (b - a) + a \tag{2.5}$$

2.2.3 Performance metrics

In supervised learning, it is important to define a single metric for evaluating the predictive performance of a learning algorithm. This allows different algorithms and hyperparameters to be methodically compared and the best option selected. In binary

classification with a negative and positive class (0 and 1, respectively), four predictive outcomes are possible: true positives (TP), false positives (FP), true negative (TN), and false negatives (FN). These outcomes are often represented in a tabular form, called the *confusion matrix*.[106]

Table 2.1 Confusion matrix of a binary classification problem

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

The following performance metrics may be calculated from the classification outcomes:

1. **Accuracy**: The fraction of total correctly predicted outcomes

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{2.6}$$

2. **Sensitivity:** The fraction of positives correctly predicted. Also known as recall or true positive rate.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.7}$$

3. **Specificity**: The fraction of negatives correctly predicted. Also known as selectivity or true negative rate.

$$Specificity = \frac{TN}{TN + FP} \tag{2.8}$$

4. **Precision**: The fraction of predicted positives that are actual positives. Also known as positive predictive value

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

5. **Negative predictive value (NPV)**: The fraction of predicted negatives that are actual negatives.

$$NPV = \frac{TN}{TN + FN} \tag{2.10}$$

6. **F1 score**: The harmonic mean of recall and precision. The F1 score combines both recall and precision into a single metric so that classifiers with a high F1 score have a high recall, i.e., correctly predict a high fraction of positives, and a high precision, i.e., do not predict many negatives as positives.

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2.11}$$

7. **F-beta score**: a beta parameter allows one to assign more weight to either the precision or recall in the F1 score. If beta is 1, the F-beta score is equivalent to the F1 score and precision and recall have equal weights. Smaller beta-values give more weight to precision, and larger values give more weight to recall.

$$F_{beta} = \frac{(1 + \beta^2) \times precision \times recall}{(\beta^2 \times precision) + recall} \tag{2.12}$$

8. **Matthew's correlation coefficient (MCC)**: All the above metrics suffer from a common problem. If the dataset is unbalanced, or if the classifier is strongly biased and tends to place most values in positive or negative class, these metrics may fail to capture that something is wrong with the classification. MCC values range from -1 through 0 to +1. A value of zero indicates that the classifier does not perform better than random guessing, whereas +1 indicates a perfectly accurate classifier, and -1 a perfectly inverse classifier. The MCC is a much-preferred metric for evaluating the overall performance of the classifier, since it considers all elements of confusion matrix.[106, 109]

$$MCC = \frac{(TN \times TP) - (FP \times FN)}{\sqrt{(TN+FP) \times (TN+FN) \times (TP \times FP) + (TN \times FN)}} \quad (2.13)$$

Alternatively, some researchers use the receiver operating curve (ROC) for an unbiased overall estimate of classifier performance. The ROC is a plot of the true positive rate (sensitivity) on the y-axis against the false positive rate (i.e., 1 – specificity). A straight line at 45 degrees to the horizontal axis indicates random guessing and a curve above this line indicates a performance better than random. The area under the curve (AUC) is used to summarize the performance, and takes values ranging from 0.5 (random guess) to 1.0 (perfect prediction). Some studies have raised concerns about the use of ROC AUC for evaluating model performance. These concerns include the large differences between true and estimated metrics for small datasets,[110] the fact that the AUC estimation is derived from ROC space in which one would rarely operate,[111] and the fact the AUC implies different misclassification cost distributions for different classifiers.[112] In this work, we consistently use the MCC for evaluating binary classifiers.

Regression performance metrics basically compare the continuous predicted and target variables. Popular metrics include the mean squared error (MSE), the root mean squared error (RMSE), the mean absolute error (MAE), cosine similarity, Pearson's correlation coefficient (R), coefficient of determination (R^2), and mean percentage error (MAPE). In this work, we use the MSE and R^2 to evaluate regression performance.

2.2.4 Dealing with data imbalance

The distribution of the dataset used in machine learning can significantly compromise the performance. Machine learning algorithms are based on the inherent assumption that the data are balanced or uniformly distributed. When the data are

imbalanced or skewed, it causes the algorithm to favor the prediction of the more abundant target values in the data at the expense of the sparse data.[113] Hence, it is imperative that the distribution of the data be altered and balanced before feeding it to a machine learning algorithm. There are broadly two types of resampling techniques in classification to abate an imbalanced distribution: undersampling and oversampling. In undersampling, the majority class is randomly reduced to achieve a balanced ratio with the minority class. In oversampling, random samples from the minority class are duplicated to achieve a balanced ratio with the minority class.[114] There are many other resampling methods that are an extension of random oversampling and random undersampling, and that often yield much better performance,[115] but the most widely used method is the synthetic minority oversampling technique (SMOTE).[116, 117] SMOTE performs both undersampling of the majority class and oversampling of the minority class. However, rather than duplication samples in the oversampling step, synthetic samples are generated by randomly interpolating between existing samples.

Data imbalance in regression problems occurs as a non-uniform distribution of the continuous target variable.[118] Compared with classification, fewer strategies have been proposed for dealing with imbalance in regression.[114, 115] It is surprising that many studies do not deal with the non-uniform distribution in supervised-learning regression problems, resulting in a sub-optimal performance at regions with sparse data.[118] In supervised-learning regression problems in this work, we implement resampling strategies that mitigate the non-uniform distribution of the target values, such as SMOTER, SMOGN, WERCS, and REBAGG.[115, 118-122] A detailed description of these methods is presented in Chapter 7.

2.2.5 Supervised learning methods

2.2.5.1 Logistic regression

The logistic regression algorithm is a linear model for categorical target values that aims to estimate the likelihood that a combination of independent variables (features) result in a certain class.[123] This estimation is achieved by using a logistic function on top of a linear combination of the independent variables. Where *X* represents a feature vector of an instance of the data to be classified, with *n* features, x^1, x^2,… x^n, and w^1, w^2,… w^n represent corresponding weights of the linear function with the bias term, w^0, the simplest case of the logistic regression model is described by equations 2.14 and 2.15 below.[96]

$$z = w^0 + \sum_{i=1}^{n} w^i x^i \tag{2.14}$$

$$\hat{y} = p(C = 1|X) = \frac{1}{1 + e^{-z}} \tag{2.15}$$

The logistic regression model is fitted to the training dataset by an iterative technique, such as Newton-Raphson, to derive optimal values for the *n*+1 weights. To prevent overfitting and improve regularization, a penalty term is often added to equation 2.12 (such as the L1 or L2 norm of the weight vector, *W*).

2.2.5.2 K-nearest neighbor

The KNN algorithm is a simple, yet powerful classification and regression technique. The algorithm is based on the idea that the target of an instance depends on the labels of the nearest neighbors in the feature space.[124] The nearest neighbors may be determined using several distance methods, such as Euclidean distance, Manhattan distance, cosine distance etc. Consequently, the predicted target is the mode of the target

values of the nearest k-neighbors, in a classification problem, or the mean of the target values, in a regression problem.

2.2.5.3 Support vector machine

The SVM algorithm seeks to identify a hyperplane in a projected non-linear mapping of the feature space that optimally separates the data into categories corresponding to the target labels.[125] The margin is the positive distance between the decision hyperplane and the instances closest to it. The goal of the SVM is to find a hyperplane that maximizes the margin, as it is expected that the larger the margin the better the generalization of the classifier.[96, 126]

2.2.5.4 Decision trees

Decision trees classify through a series of questions, beginning from a root node at the top, in which the next question is dependent on the answer to the last question, and each step in the series results in purer dataset as you progress through other nodes down the tree until you reach terminal or leaf nodes.[96, 127, 128] The quality of the split at each node is measured by an impurity score, and a purer split results if the proportion of a class increases in a branch after the split. At a given node, *m*, the probability that an instance is in a class C_i of *K* classes is given by the ratio of the number of instances in C_i to the total number of instances at node *m*, as described by equation 2.16. Impurity is popular measured via entropy (equation 2.17) or the Gini index (equation 2.18).[127, 129]

$$p_m^i(C_i|x,m) = \frac{N_m^i}{N_m} \tag{2.16}$$

$$I_m = -\sum_{i=1}^{k} p_m^i \, log(p_m^i) \tag{2.17}$$

$$I_m = 1 - \sum_{i=1}^{k} (p_m^i)^2 \tag{2.18}$$

The decision tree may recursively split each node until perfectly pure leaf nodes are obtained, but this may lead to overfitting, so a stopping criterion may be defined. The result of each split of the decision tree is the partitioning of the feature space into distinct rectangular regions which allows for a non-linear and non-parametric function mapping the features to target values.

2.2.5.5 Random forest

A random forest, as the term suggest, comprises of many trees combined to form an bootstrap aggregate ensemble, or a *bagging ensemble*.[130, 131] While decision trees may be unstable and tend to overfit, a combination of many trees on bootstrap samples of the dataset, such that the final outcome is the aggregate of the outcomes of individual trees (mode for classification, mean for regression), results in a stable model with markedly improved generalization due to reduced variance.[132] Random forest has achieved an overwhelming level of success in biological machine learning problems particularly for the following reasons:[132-136]

1. It is nonparametric and makes no assumptions on the distribution of the data, allowing for complex data structures.
2. It is computationally efficient.
3. It robustly tolerant to noise in the dataset.

4. It can handle a small sample size and high-dimensional feature space as it implements an implicit feature selection in the learning process.
5. It is an interpretable model that allows the user to gain deeper insight into the data via feature importance measures.

The importance of a feature in the random forest model can be evaluated by a quantitative measure, such as the Gini importance, to provide knowledge on the relative relevance of each feature in the predictive process. The Gini importance of a feature is calculated as the average as the average decrease in the Gini impurity (equation 2.18) at every split in the in the forest where the feature was used as the splitting variable. Feature importance provides an efficient and highly practical quantitative basis for develop hypotheses that relate biological features to a specific attribute. For example, important residues for enzyme catalytic activity and potential targets for protein engineering may be identified by determining what residues yield the highest feature importance in a random forest prediction of activity.

2.3 Protein conservation analysis

A central paradigm in protein science is that protein sequences sharing common evolutionary history can be aligned such that similar positions that are a result of functional, structural, and evolutionary relationships can be identified.[137] Several algorithms exist for aligning multiple homologous protein sequences such that all sequences in the multiple sequence alignment (MSA) have the same length. Generally, MSA algorithms seek to optimize a scoring (objective) function, such as PAM or BLOSUM substitution matrix, that defines the cost of substituting one amino acid for another in the alignment. A gap

penalty is used to estimate the cost of introducing indels in the alignment. By heuristically minimizing the total cost, usually using a pre-computed guide tree, the MSA algorithm achieves a resulting alignment that is a decent trade-off between structural accuracy and computational efficiency.[138] Widely used MSA algorithms include ClustalW,[139] Clustal Omega,[140] T-Coffee,[141] ProbCons,[142] MUSCLE,[143] and MAFFT.[144]

From an MSA, a set of homologous protein sequences (rows in the alignment) with aligned amino-acid residues (columns) can be examined to gain insight to the conservation patterns of evolutionarily similar residue positions. The fundamental basis of conservation analysis is the idea that positions in the MSA that are under significant evolutionary pressure for functional or structural roles have different amino-acid distributions and are more conserved than other less relevant positions.[145] Consequently, determining what positions are conserved can provide an evolutionary basis for determining sites in proteins that play key roles in functional variation. Despite its statistical simplicity, conservation analysis is considered to be the single most powerful predictor of functionally relevant sites in proteins,[145, 146] and has been notably successful in identifying residues that play key roles in ligand binding,[147, 148] protein-protein interaction,[149-151] functional specificity,[152-154] and structural stability.[155-157] It follows that conservation analysis is a powerful tool for identifying promising sites for protein engineering.

There are diverse metrics for evaluating the conservation of a position in an MSA, and each metric has unique advantages and disadvantages.[158] Given that p_i^x is the probability (or frequency) that amino acid, *x*, occurs at site, *i*, in the MSA, and p_{MSA}^x is the probability that amino acid, *x*, occurs in the MSA, below are some popular conservation evaluation metrics:

1. **Shannon entropy**: the simplest conservation measure, but does not take into consideration the background distribution of amino-acids in the MSA or the similarity between amino acids.

$$SE_i = \sum_x -p_i^x \log\left(p_i^x\right) \tag{2.19}$$

2. **Relative entropy (Kullback-Leibler divergence)**: an improvement over Shannon entropy as it considers the background distribution, but does not consider the similarity between amino acids.

$$RE_i = \sum_x p_i^x \log\left(\frac{p_i^x}{p_i^{MSA}}\right) \tag{2.20}$$

3. **Lockless evolutionary conservation parameter**: a conservation metric similar to the relative entropy put forward by Lockless and Ranganathan.[159]

$$\Delta G_i = \sqrt{\sum_x \left(ln \frac{p_i^x}{p_{MSA}^x}\right)^2} \tag{2.21}$$

Other metrics include the Jensen-Shannon divergence score,[160] Rate4Site,[161] Schneider score,[162] Kabat score,[163] Landgraf metric,[164] and Real Evolutionary Trace (RET).[165] In this work, we use the relative entropy score because of its simplicity and because frequency-based conservation scores have been observed to outperform others.[158] While there are available web-servers and command-line software for performing conservation analysis, such as Consurf,[166] Scop3D,[167] and BALCONY,[168] we prepared an open-source Python package for conservation analysis, PyCanal, that allows conservation analysis to be implemented within the Python framework, taking advantage of the valuable BioPython library.[169] PyCanal is available at https://github.com/jafetgado/PyCanal.

2.4 Phylogenetic analysis

Phylogenetics is the study of the evolutionary relationships among organisms or biological products, such as genes and proteins. Phylogenetics plays a major role in the scientific understanding and systematic classification of species.[170-172] Phylogenetic analysis of proteins seeks to reconstruct the evolutionary history, which provides insight into functional and structural diversity of related proteins. The output of phylogenetic analysis is a phylogenetic tree, which shows a hypothesis of the evolutionary history and has a unique branching pattern (the topology). Taxa that are closed together in the tree are more closely related and share a common ancestor. The length of the tree branches represents the evolutionary time between nodes and is in unit of number of substitutions per sequence site.

For *m* taxa (or products), the number of possible rooted bifurcating tree topologies is $\frac{(2m-3)!}{2^{m-2}(m-2)!}$, so that there are over 34 million possible topologies for only 10 products.[172, 173] As a result, heuristics and optimization techniques are employed in inference methods to find the best topology that fits the sequence data.[174] There are generally three types of phylogenetic tree-building methods: parsimony, likelihood, and distance methods. Likelihood methods seek to find the tree topology that achieves the highest probability of observing the sequence data for a specific substitution model. Parsimony methods seek to find the topology that requires the fewest substitutions or changes. They are based on the idea that the best tree is the simplest tree with minimal assumptions.[175] Distance methods use evolutionary models to evaluate the pairwise distance between the sequences, i.e. the number of substitutions between sequences, and then infer the topology from the computed distances.[176]

The evolutionary distance between protein sequences may be estimated from the MSA by mathematical models such as the gamma function,[177] or the Grishin distance,[178] or by a substitution matrix such as the Dayhoff matrix,[179] or the Jones et al (JTT) matrix.[180] Substitution matrices present the probability an amino acid in the i^{th} row will change to the amino acid in the j^{th} column during a defined evolutionary time unit. Having selected a model for computing pairwise evolutionary distances, the phylogenetic tree can be inferred by a number of distance methods such as the unweighted pair-group method with arithmetic means method (UPGMA), the least squares method (LS), and minimum evolution method. Minimum evolution method is widely used due its simplicity and a high accuracy that is comparable to other more computationally expensive methods.[181] The minimum evolution method selects the topology with the minimum sum of all branch lengths (distances) as the optimal topology. The computational cost of the minimum evolution method is significantly reduced by using the close-neighbor interchange algorithm (CNI).[182] CNI searches for the optimal tree by quickly computing a neighbor-joining tree and then iteratively searching for topologies that are a few permutations from the neighbor-joining tree and which yield a smaller total distance value than the neighbor-joining tree.

Confidence values for each node in the phylogenetic tree may be estimated using a bootstrap test.[183] In the bootstrap test, *n* sites are randomly chosen with replacement to be reshuffled, and a new phylogenetic tree is inferred from the reshuffled alignment. The topology of the new tree is compared to the original tree and interior branches that yield the same partition of sequences in both the original and reshuffled tree are given a score of 1, and, otherwise, a score of 0. This process is repeated numerous times (usually about 500

to 1000 times), and the percent of times each branch receives a value of 1 is computed as a percent. This value is the bootstrap confidence value and is a measure of the confidence or accuracy of the inferred topology.[172, 184]

CHAPTER 3. Machine Learning Reveals Sequence-Function Relationships in Family 7 Glycoside Hydrolases

In this chapter, we applied supervised learning and bioinformatic analysis to investigate the relationships between amino-acid sequence and functional variation in family 7 glycoside hydrolases. The author of this dissertation performed all computational experiments in this chapter. The director of this dissertation (Christina M. Payne) and a collaborator at Department of Computer Science (Brent Harrison) helped with the experiment design. Collaborators at Swedish University of Agricultural Sciences (Mats Sandgren and Jerry Ståhlberg) provided assistance in the manual curation of sequence alignments.

3.1 Abstract

Family 7 glycoside hydrolases (GH7) are among the principal enzymes for cellulose degradation in nature and industrially. These important enzymes are often bimodular, comprised of a catalytic domain attached to a carbohydrate binding module (CBM) via a flexible linker, and exhibit a long active site that binds cello-oligomers of up to ten glucosyl moieties. GH7 cellulases consist of two major subtypes: cellobiohydrolases (CBH) and endoglucanases (EG). Despite the critical biological and industrial importance of GH7 enzymes, there remain gaps in our understanding of how GH7 sequence and structure relate to function. Here, we employed machine learning to gain insights into relationships between sequence, structure, and function across the GH7 family. Machine-learning models, using the number of residues in the active-site loops as features, were able

discriminate GH7 CBHs and EGs with up to 99% accuracy. The lengths of the A4, B2, B3, and B4 loops were strongly correlated with functional subtype across the GH7 family. Position-specific classification rules were derived such that specific amino acids at 42 different sequence positions predicted the functional subtype with accuracies greater than 87%. A random forest model trained on residues of 19 positions in the catalytic domain predicted the presence of a CBM with 89.5% accuracy. We propose these positions play vital roles in the functional variation of GH7 cellulases. Taken together, our results complement numerous experimental findings and present functional relationships that can be applied when prospecting GH7 cellulases from nature, for sequence annotation, and to understand or manipulate function.

3.2 Introduction

Cellulose is the most abundant renewable biopolymer on Earth and, thus, holds tremendous potential in transitioning energy production from fossil fuels to a renewable carbon feedstock — a key need to limit anthropogenic climate change. Sugars derived from the deconstruction of cellulose can be converted to biofuels and numerous chemicals via myriad biological or catalytic conversion routes. However, the efficient depolymerization of cellulose in a cost-effective manner such that biofuels can economically compete with fossil fuels remains a major challenge to enabling a lignocellulosic economy.[10] In industry, biochemical methods of cellulose deconstruction employing enzymes are promising due to high selectivity, low energy consumption, and low amounts of by-product generation.[10, 185, 186] As a result, improving the yield of enzymatic hydrolysis of cellulose by enhancing cellulase activity is a major research focus.

In nature, microbial cellulose degradation is primarily achieved via a synergistic cocktail of enzymes consisting of processive cellobiohydrolases (CBHs), endoglucanases (EGs), and accessory enzymes such as β-glucosidases and lytic polysaccharide monooxygenases (LPMOs).[185] Organisms can employ these enzymes as free single- or multi-modular constructs, or as cellulosomes. Industry tends to employ free enzyme systems, as filamentous fungal hosts are proficient secretors of these types of cellulose-degrading enzymes. EGs act by attacking internal bonds in cellulose, thus, creating free chain ends. CBHs attach to free chain ends via exo-initiation, or internal regions in the chain via endo-initiation, and processively cleave off cellobiose units as they process along the chain. Cellobiose products are consequently hydrolyzed by β-glucosidases to yield glucose.[185]

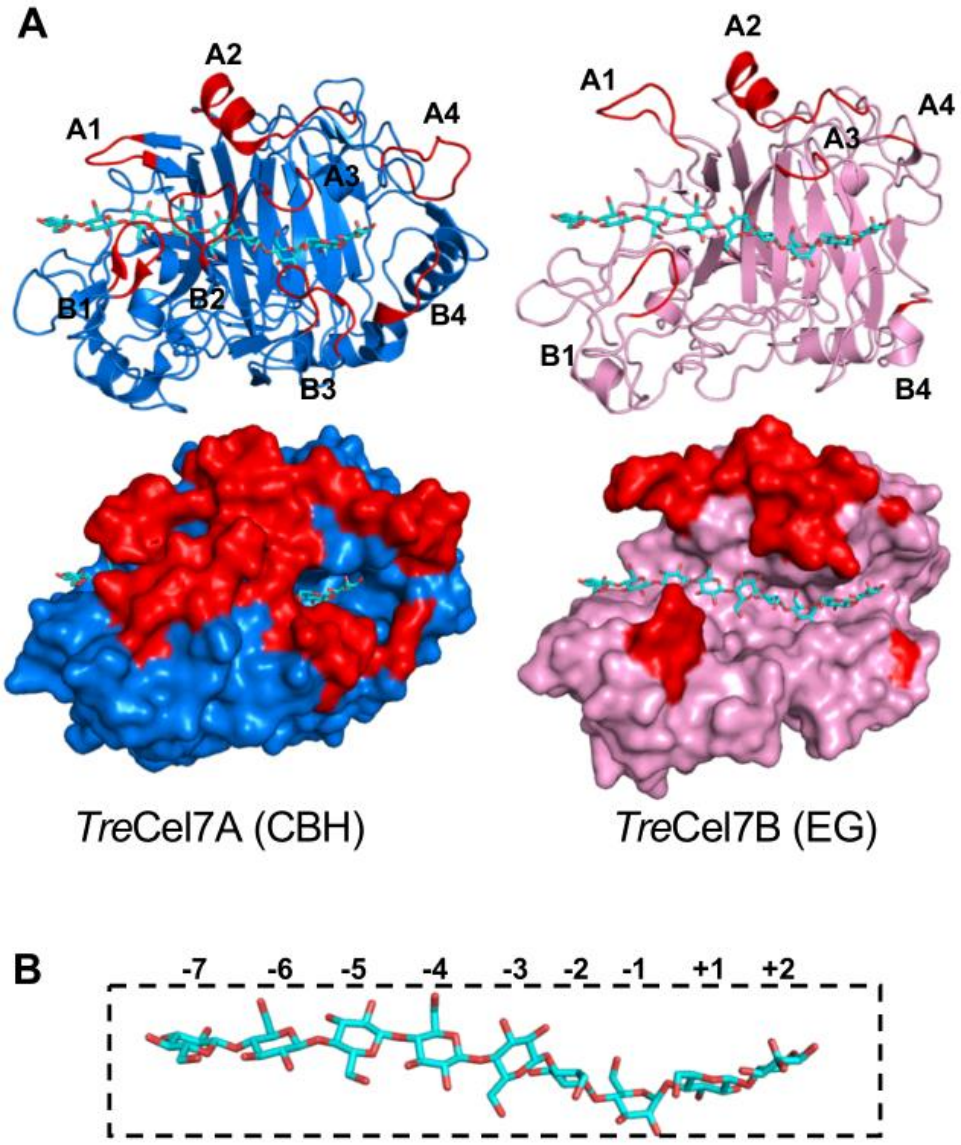


Figure 3.1 Structures of typical GH7 CBH and EG with a cellononaose ligand in complex. (A) The CBH (left), *Trichoderma reesei* Cel7A (*Tre*Cel7A, PDB code: 4C4C),[187] and the EG (right), *Trichoderma reesei* Cel7B (*Tre*Cel7B, PDB code: 1EG1).[188] The eight active-site loops (A1 to A4 and B1 to B4) are shown in red. In the CBH, the active site is tunnel-like, but is more open and groove-like in the EG. (B) Glycosyl binding sites are numbered from the non-reducing end at the active-site tunnel entrance (-7) to the reducing end (+2)

where the cellobiose product exits the active site. Bond cleavage occurs between -1 and +1 subsites.

Whereas CBHs are known to be processive and to carry out several cellulolytic cuts before detaching from the cellulose substrate, EGs are mostly nonprocessive or may show little processivity.[41, 189-192] Optimum cellulolytic efficiency is achieved by the synergistic action of CBHs and EGs. CBHs, EGs, and β-glucosidases, as well as other glycoside hydrolases (GHs) are currently classified into 168 families in the CAZy database.[31, 193] Family 7 glycoside hydrolases (GH7s) are the powerhouses of cellulose degradation in nature. They traditionally are found mostly in fungi, although sequences have been identified in several non-fungal groups such as Crustacea, Porifera, Alveolata, and Amoeba.[42] Because GH7s offer significant cellulolytic potential, they are often the predominant enzymes by mass in the secretomes of many filamentous cellulolytic fungi and constitute the major components of enzyme cocktails in industrial cellulolytic processes.[185, 194, 195]

GH7s consist of two main subtypes, CBHs and EGs. Although over 5,000 GH7 sequences are known, structural information is presently available for only 21 GH7s (16 CBHs, 5 EGs).[35, 36, 39, 42, 46, 187, 188, 196-207]. GH7 CBH and EG structures share a similar β-jelly roll fold with two antiparallel β-sheets that pack into a curved β-sandwich.[36] Loops protrude from the β-sandwich and extend over a tunnel-like active site that spans 40–50 Å across the ends of the catalytic domain (CD). The active site contains at least nine glycosyl subsites for binding cello-oligomers, which are numbered -7 to +2 from the non-reducing end of the cellulose chain (Figure 3.1). The cellulose chain is cleaved between the -1 and +1 subsites.[185] Despite the overall similarity in fold, structures of GH7 CBHs and EGs are

strikingly different in their active-site configuration. Whereas GH7 CBHs exhibit a closed tunnel-like active site, GH7 EGs possess a more open, groove-like active site. These differences arise due to the variation in the residue lengths of the loops that protrude over the active-site groove, labeled A1 to A4 and B1 to B4 (Figure 3.1).[35] Several structural and mechanistic studies of GH7s have proposed that the differences in functional properties of GH7 CBHs and EGs, such as processivity, endo-initiation, and product inhibition, arise mainly due to the differences in the active-site architecture in the loops.[35, 39, 197, 200, 201, 203, 204, 206] Moreover, GH7 CBHs with a more exposed active-site tend to exhibit functional characteristics intermediate between typical CBH and EG behavior.[40, 41, 200, 208] Besides the differences in the configuration of active-site loops, studies have also indicated that there are key residues in the active site of GH7s that contribute to the variation in GH7 CBH and EG behavior. Several aromatic and charged residues in the active site that interact with the cellulose substrate have been suggested to be crucial for the processive activity of GH7 CBHs.[45, 202, 209-211] Moreover, mutation of these residues notably diminishes the processive activity of GH7 CBHs on crystalline cellulose.[212, 213]

Like many other cellulases, GH7s can be bimodular, having their CD attached to a carbohydrate binding module (CBM) by an intrinsically disordered glycosylated linker peptide.[37, 214-217] There are currently 87 families of CBMs in the CAZy database.[193, 218] but GH7s mainly utilize family 1 CBMs.[185, 219] It is now generally accepted that family 1 CBMs function to increase the affinity of cellulases for crystalline cellulose and, thereby, increase the surface concentration of the enzyme for catalysis. Thus, by facilitating two-dimensional diffusion of the CD on the cellulose surface, the CBM improves the catalytic efficiency.[37] Furthermore, several studies have revealed that deletion of the CBM-linker domain

dramatically reduces CBH activity on crystalline cellulose, especially at low enzyme concentration, but not on soluble substrates.[38, 219-223] Takashima *et al*. carried out several mutations in the CBM of a *Humicola grisea* CBH (*Hgr*Cel7A) and observed high positive correlation between the efficiency of the enzyme on crystalline cellulose and the binding affinity of the CBM.[224] Similarly, Srisodsuk *et al*. observed that replacing the CBM of *Trichoderma reesei* Cel7A (*Tre*Cel7A) with the CBM of *Tre*Cel7B, which has a higher cellulose-binding affinity, improved the activity of *Tre*Cel7A on crystalline cellulose.[38] Altogether, these results indicate that CBMs affect GH7 catalytic activity primarily by promoting binding to the cellulose surface.

Despite the tremendous growth in scientific knowledge of GH7s over the last few decades, our understanding of how sequence and structure affect function is far from complete. Although it is known that the exposure of the active site due to truncation in the active-site loops can substantially affect function, little work has been done to elucidate the unique roles that each of the active site loops play and how the effects of truncation vary with function for the different loops. Recently, Schiano-di-Cola *et al.* studied the effects of deletions in the B2, B3, and B4 loops on the activity and kinetics of *Tre*Cel7A.[225] They found that deletions in the B2 loop, compared to the B3 and B4 loop, most significantly affect CBH behavior of *Tre*Cel7A. Beyond *Tre*Cel7A, there is a need to investigate how variation of active-site loop lengths relate to function across other members of the GH7 family.

In this work, we employ machine learning (ML) to derive relationships between sequence, structure, and function of GH7s using a dataset of 1,748 selected protein sequences. The sequences are aligned via multiple sequence alignment (MSA) to identify

regions of structural similarity and evolutionary importance. Although manual inspection of the MSA may reveal several functional patterns, such as highly conserved positions, many important but complex relationships may be missed. ML is an especially useful statistical tool when data are abundant and relationships in the data are complex.[226] Thus, ML can be employed to discover complex functional and evolutionary relationships in proteins. In this work, we apply ML to the MSA of GH7 sequences, mapping variation in lengths of the active-site loops to functional subtypes such that the subtype can be accurately predicted from loop length. We also derive position-specific classification rules to highlight positions that play important roles in CBH/EG function. Lastly, we investigate relationships between the CBM and the CD by utilizing ML to predict the presence of CBMs in GH7s using residues in the CD. It is important to note that, as the current understanding of GH7 function is based on investigation of a few representatives, this present study of 1,748 GH7 sequences seeks to identify general sequence-function relationships for the entirety of the GH7 family and the degree to which variation exists.

3.3 Results

3.3.1 Datasets

Three datasets were used in this study. The first dataset contained 1,748 full-length GH7 protein sequences retrieved from the National Center for Biotechnology Information (NCBI) non-redundant database. Using a strict keyword search, we queried the NCBI database for the subtype annotation (i.e. CBH or EG) of these 1,748 sequences. 427 sequences were clearly annotated as CBH or EG in the database (291 CBHs and 136 EGs), and these 427 sequences comprised the second dataset. For the third dataset, we retrieved

44 GH7 sequences from the manually curated UniProtKB/Swiss-Prot database.[227] Accordingly, the subtype annotations of the 44 GH7s (30 CBHs, 14 EGs) are less likely to contain errors than the annotations of the 427 sequences from the NCBI non-redundant database.

3.3.2 Discrimination of GH7 subtypes with hidden Markov models

In the annotation of a protein sequence, several computational prediction methods may be applied. Sequence similarity methods compare an unclassified protein with well-studied proteins and assign the unclassified protein to the same class as the most similar classified proteins.[228] Hidden Markov model (HMM),[229, 230] which describes the protein sequence as a probabilistic model, is one of the most sensitive and most accurate methods for discriminating protein functional families with sequence data alone, provided they are built with correct alignments.[228] Within a given protein family, HMM can also be applied to discriminate functional subtypes, although the discrimination accuracy varies across different families.[152]

We applied HMM to discriminate GH7 CBHs and EGs. The performance of HMM was evaluated by a five-fold cross-validation technique using the datasets of 427 (NCBI) and 44 (UniProtKB/Swiss-Prot) GH7 sequences. First, each dataset was aligned and separated into CBH and EG subalignments based on the database annotations. Then, each subalignment was randomly split into five folds (Figure 3.2A). Subtype HMMs (i.e., CBH HMM and EG HMM) were repeatedly built on four out of five folds of the CBH and EG subalignment, and the sequences in each left-out fold were used as a test set. To predict the subtype of a sequence, the sequence was aligned separately to both the CBH and EG HMMs, and then the alignment scores were compared. If the CBH HMM alignment score

was greater than the EG HMM alignment score, the sequence was predicted to be a CBH; otherwise, it was predicted to be an EG.[152] The process was repeated so that all five folds were used in training and testing the HMMs.

Figures 3.2B and 3.2C show the performance of the HMM method on the UniProtKB/Swiss-Prot dataset (44 sequences) and on the NCBI dataset (427 sequences), respectively. The HMM method achieved perfect accuracy on the UniProtKB-Swiss-Prot dataset. All sequences were correctly predicted, and there was a substantial difference, of at least 120.0, between the CBH alignment score and the EG alignment score. On the NCBI dataset of 427 sequences, which may contain erroneous subtype annotations, the HMM achieved an accuracy of 99.53% and only misclassified two sequences (accession codes: AGY80096.1 and AGY80097.1), which are annotated as EGs. These two sequences may have been erroneously annotated as EGs since they are much more similar to CBHs in overall sequence and loop lengths. Furthermore, the value of the alignment score difference for some sequences in the NCBI dataset is as low as 2.0.

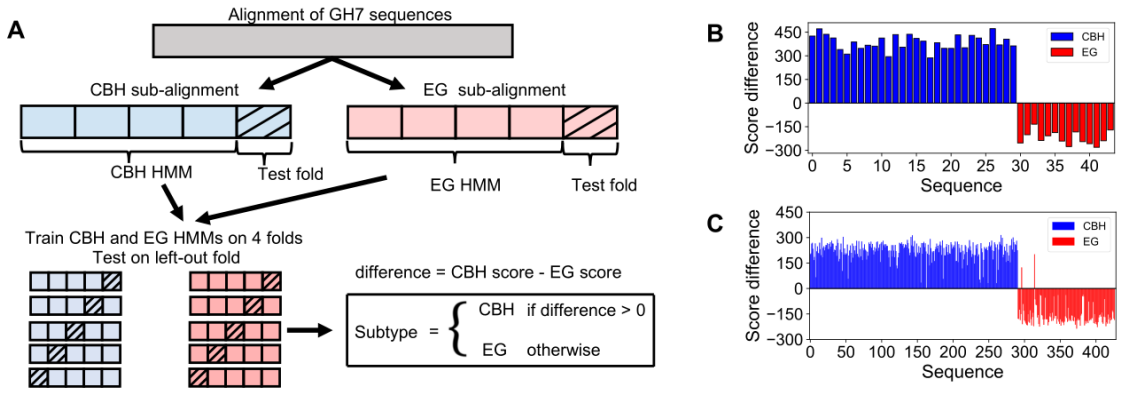


Figure 3.2 Discrimination of GH7 CBHs and EGs with hidden Markov model (HMM).(A) Five-fold cross-validation technique for evaluating the performance of HMM. The MSA is split into CBH and EG subalignments and each subalignment into five folds. HMMs are

repeatedly trained on four folds and then tested on the left-out fold. The predicted class (CBH or EG) of a sequence is the class that yields the highest HMM alignment score. (B) Performance of HMM on the dataset of 44 GH7s from the manually curated UniProtKB/SwissProt database. (C) Performance of HMM on the dataset of 427 GH7s from NCBI non-redundant database. Only two EG sequences (GenBank accession codes: AGY80096.1 and AGY80097.1) were misclassified in the NCBI dataset. Note that in B and C, the assigned sequence numbers (*x*-axes) are arbitrary.

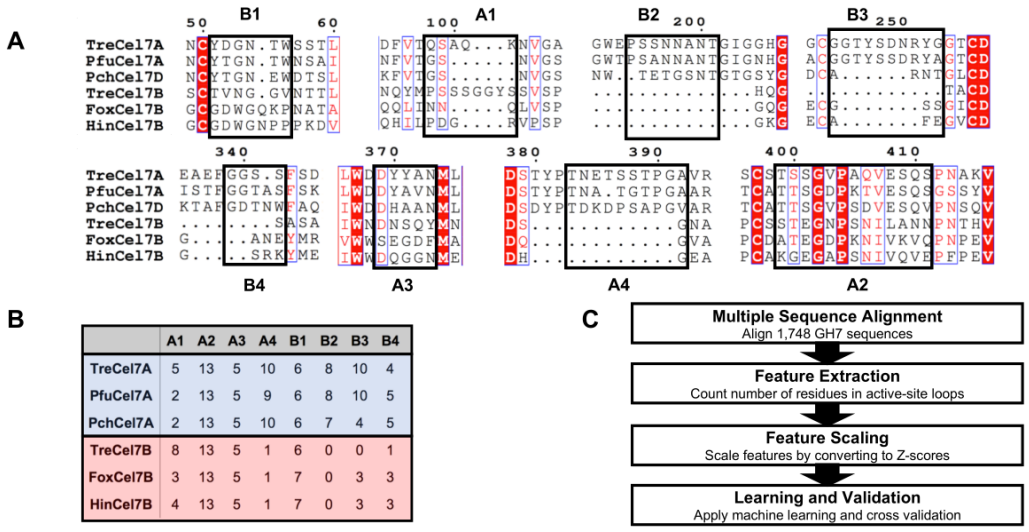
3.3.3 Discrimination of GH7 subtypes with machine learning: relationships between active-site loops and CBH/EG function

In this part of the study, our goal was to use ML to map the variation in amino acid sequence to GH7 CBH and EG activity and to, consequently, determine which aspects of the sequence and structure predominantly affect CBH/EG function. If a particular feature is important for the difference in CBH and EG behavior, we should be able to train ML models on that feature to discriminate GH7 CBHs and EGs with significant accuracy. Otherwise, a feature that has no correlation with activity, but only varies due to phylogenetic diversity, would perform poorly when applied to predict GH7 subtypes with ML.

We used the dataset of 1,748 GH7s to test ML algorithms in predicting GH7 subtypes. Since only 427 of the 1,748 GH7s are classified as CBH or EG in the databases, we applied the HMM method described previously to derive the functional classes of the unclassified GH7 sequences. Our cross-validation tests showed that the HMM method can correctly classify GH7 subtypes with an accuracy of almost 100% (i.e. consistent with the

database annotations). This result is similar to the performance of the HMM method applied to other protein families.[152] Moreover, when we trained separate HMMs on the manually-annotated dataset of 44 sequences (UniProtKB/Swiss-Prot) and on the "less perfect" dataset of 427 sequences (NCBI), and then applied the HMMs to determine the subtype of the 1,748 GH7s, the separate HMMs assigned the same subtype in all but five instances (99.71%). Regardless, misclassification errors of about 1% are not large enough to alter the relationships that we derived from ML on the dataset of 1,748 GH7 sequences.[231, 232]

In choosing features for the ML models, we capitalized on the observation that crystal structures of GH7 CBHs and EGs differ in their active-site architecture, due to the degree of truncation in the eight active-site loops (Figure 3.1). Hence, we used the number of residues in the active-site loops as features for ML to discriminate between GH7 CBHs and EGs. First, a structure-based MSA of all 1,748 sequences was carried out (See Materials and Methods for details). For each sequence in the MSA, we counted the number of amino acid residues in the eight active-site loops and derived a vector of the eight loop lengths as features (Figure 3.3).



	A1	A2	A3	A4	B1	B2	B3	B4
TreCel7A	5	13	5	10	6	8	10	4
PfuCel7A	2	13	5	9	6	8	10	5
PchCel7A	2	13	5	10	6	7	4	5
TreCel7B	8	13	5	1	6	0	0	1
FoxCel7B	3	13	5	1	7	0	3	3
HinCel7B	4	13	5	1	7	0	3	3

Figure 3.3 Generating features for discriminating GH7 CBHs and EGs with machine learning. (A) Segments of a selection of six well-studied GH7s from the structure-based sequence alignment of 1,748 sequences showing the active-site loops. The sequences include the CBHs: *Trichoderma reesei* Cel7A (*Tre*Cel7A),[187] *Penicillium funiculosum* Cel7A (*Pfu*Cel7A),[46] and *Phanerochaete chrysosporium* Cel7D (*Pch*Cel7D);[200] and the EGs: *Trichoderma reesei* Cel7B (*Tre*Cel7B),[188] *Fusarium oxysporum* Cel7B (*Fox*Cel7B),[203] and *Humicola insolens* Cel7B (*Hin*Cel7B).[204] (B) The number of residues in the eight active-site loops as determined from the structure-based alignment. (C) Procedure for generating features for 1,748 GH7s. First, the sequences are aligned as in (A). Then, a count of the number of residues in each loop is obtained. Residue counts are scaled to Z-scores before ML is applied.

Four ML methods were applied: decision trees, logistic regression, k-nearest neighbors (KNN), and support vector machines (SVM). For each ML method, nine models with different combinations of features were tested. One model involved training the ML algorithms on the lengths of all eight loops, and the remaining 8 models involved using

each loop length as the sole feature for the training (single-feature models). The performance of the ML models was measured using four metrics: sensitivity (or true positive rate), specificity (or true negative rate), overall accuracy, and Matthew's correlation coefficient (MCC). Here, the sensitivity is the percent of CBHs (the true class) correctly predicted, the specificity is the percent of EGs (the false class) correctly predicted, and the overall accuracy is the percent of both CBHs and EGs correctly predicted. The MCC ranges from -1 to +1 and measures the correlation between the predicted and true classifications. An MCC value of +1 indicates perfect prediction, 0 indicates no concordance between predicted and actual classes, and -1 indicates perfect disagreement. MCC has been recommended as the most informative performance metric in evaluating binary classification performance, especially when the dataset is imbalanced since other metrics such as overall accuracy and F1 score can be hugely misleading.[106, 233-235] Hence, we use MCC as the primary metric in evaluating the performance of the ML models.

Moreover, we are faced with the problem of an imbalanced dataset: 1,306 (75%) of the 1,748 sequences in the dataset are CBHs. Ordinarily, imbalanced data will skew the results by causing the ML classifiers to place most of the data in the majority class (CBH). To deal with the imbalance problem, we applied random undersampling to the majority class so that the distribution of CBH and EGs was balanced.[236, 237] We evaluated the performance of the ML models on the redistributed data with 100 repetitions of five-fold cross validation, with the dataset undersampled and reshuffled in each repetition (Figure 3.4). Repeating the five-fold cross validation numerous times is a highly effective way to mitigate the effects of variability in the train-test splits and to ensure that the data space is thoroughly explored despite loss of data in the undersampling step.[107]

Our results show that ML is able to accurately discriminate between GH7 CBHs and EGs using only information about the length of the active-site loops (Table 3.1). However, the performance varied significantly for the different single-feature models (Figure 3.5A). The models trained on the A2 and A3 loops exhibited the worst performance with MCC values close to zero, indicating that they did not perform better than a random classification. The models trained on A1 and B1 loops showed intermediate performance with MCC values widely varying from -0.08 to 0.79 for the A1 models, and -0.03 to 0.63 for the B1 models. Interestingly, the A4, B2, B3, and B4 models showed very high predictive performance, with MCC values ranging from 0.94 to 0.98 and with much lower variation among the different ML methods. The models trained on these five loops (A4, B2, B3, B4) achieved nearly the same high performance as the models trained on all eight loops (Table 3.1) .

Table 3.1 Performance of machine learning algorithms in discriminating GH7 CBHs and EGs.The first eight rows show the performance of the models trained on each loop as the single, independent feature. The last row shows the performance of the models trained with all eight loops as features.

	Decision tree			Logistic regression			K-nearest neighbor			Support vector machine		
Features	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
A1	98.6 ± 1.2	45.9 ± 5.0	72.3 ± 3.2	42.0 ± 16.5	52.8 ± 7.3	46.9 ± 6.4	86.5 ± 15.1	88.7 ± 5.4	87.6 ± 5.6	97.0 ± 1.8	85.5 ± 3.4	91.2 ± 2.0
A2	65.9 ± 43.5	37.4 ± 42.2	50.7 ± 3.7	49.3 ± 46.7	50.3 ± 45.3	47.4 ± 2.8	4.6 ± 2.3	97.0 ± 1.9	50.8 ± 3.8	89.2 ± 27.3	18.4 ± 26.0	53.0 ± 3.5
A3	89.0 ± 26.3	16.9 ± 24.8	52.5 ± 3.7	50.8 ± 47.9	49.4 ± 45.4	47.6 ± 3.2	3.0 ± 2.0	97.8 ± 1.6	50.4 ± 3.7	96.7 ± 11.2	11.4 ± 10.9	53.9 ± 3.4
A4	95.7 ± 2.1	99.5 ± 0.7	97.6 ± 1.1	95.8 ± 2.0	99.5 ± 0.6	97.7 ± 1.1	95.8 ± 2.1	99.7 ± 0.5	97.8 ± 1.1	95.6 ± 2.2	99.6 ± 0.6	97.6 ± 1.1
B1	96.8 ± 1.8	44.1 ± 5.5	70.5 ± 3.3	79.3 ± 35.6	34.5 ± 12.2	55.9 ± 13.4	1.3 ± 1.8	98.7 ± 1.6	50.0 ± 3.7	95.1 ± 2.6	72.3 ± 4.4	83.7 ± 2.6
B2	94.6 ± 2.4	99.1 ± 1.2	96.9 ± 1.3	94.7 ± 2.4	98.4 ± 1.2	96.6 ± 1.3	95.3 ± 2.3	97.4 ± 1.8	96.4 ± 1.3	94.8 ± 2.4	98.4 ± 1.6	96.6 ± 1.4
B3	92.4 ± 2.7	99.8 ± 0.5	96.1 ± 1.4	89.9 ± 3.3	99.8 ± 0.5	94.8 ± 1.7	96.3 ± 2.2	98.6 ± 1.1	97.5 ± 1.2	89.7 ± 3.3	99.8 ± 0.4	94.8 ± 1.7
B4	97.9 ± 1.8	98.2 ± 1.3	98.0 ± 1.0	98.2 ± 1.4	98.2 ± 1.2	98.2 ± 0.9	97.6 ± 2.0	98.3 ± 1.3	98.0 ± 1.1	97.8 ± 1.6	98.2 ± 1.3	98.0 ± 1.0
All 8 loops	98.8 ± 1.2	99.1 ± 1.1	98.9 ± 0.8	98.3 ± 1.4	99.2 ± 0.9	98.8 ± 0.8	97.1 ± 2.0	99.4 ± 0.7	98.2 ± 1.1	99.0 ± 1.1	98.9 ± 1.1	98.9 ± 0.7

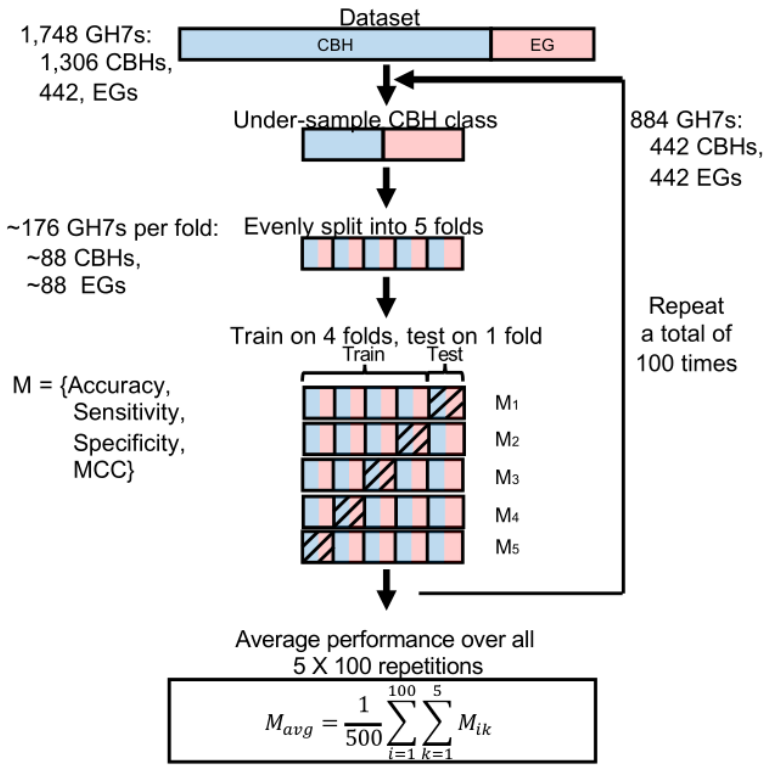


Figure 3.4 Procedure for evaluating the performance of ML models using 100 repetitions of five-fold cross validation with undersampling.The dataset is reshuffled and resampled in each repetition.

Furthermore, we observed that the variation in the lengths of the loops correlates with the discriminative performance of the loops (Figure 3.5A-C). The loops with very poor discriminatory performance (A2 and A3) show the lowest relative variation in lengths across the 1,748 GH7s, and nearly identical distributions between CBHs and EGs (Figure A1.1 of Appendix A1). In contrast, loops with intermediate discriminatory performance (A1 and B1) show a greater level of variation in lengths than A2 and A3 loops and noticeably different distributions for CBHs and EGs, although there is a considerable amount of overlap. The loops with near-perfect predictive performance (A4, B2, B3, B4) show the highest variation in lengths.

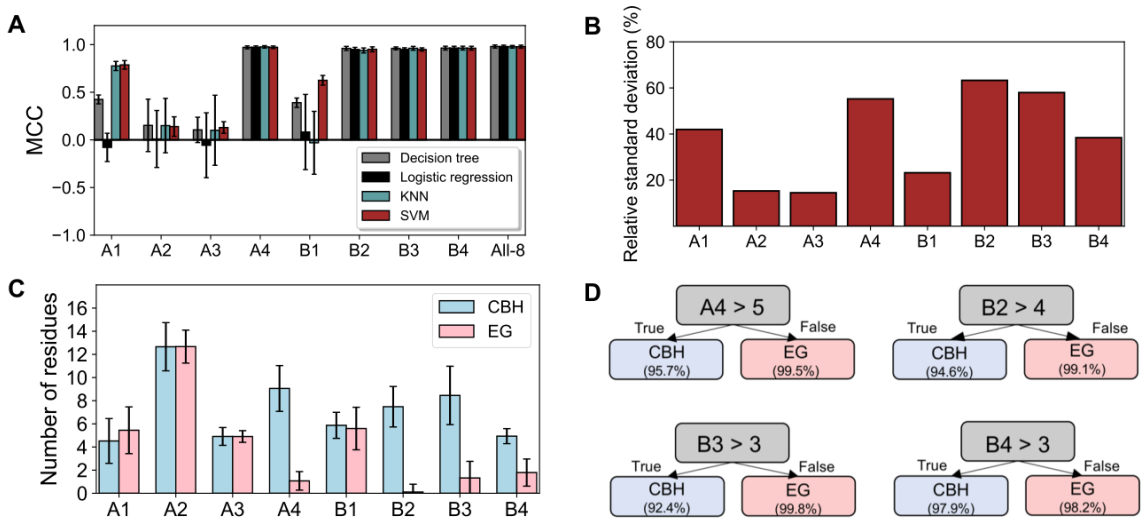


Figure 3.5 Predictive performance and variation of active-site loops in GH7s.(A) Matthews' correlation coefficient (MCC) values of four ML algorithms trained separately on the length of each active-site loop and on all eight loops together. The A4, B2, B3, and B4 loops achieve near-perfect performance in discriminating 1,748 GH7 CBHs and EGs. (B) The relative standard deviation of the length of the eight active site loops. Generally, variation in the length of a loop correlates with predictive performance of the loop as a ML feature. (C) The mean length of active-site loops in 1,306 GH7 CBHs and 442 GH7 EGs. Error bars are ± 1 standard deviation. (D) Rules derived from the single-node decision trees trained on the A4, B2, B3, and B4 loops. The accuracy of the rules in discriminating GH7 CBHs and EGs, i.e. the sensitivity and specificity, respectively, are shown in brackets.

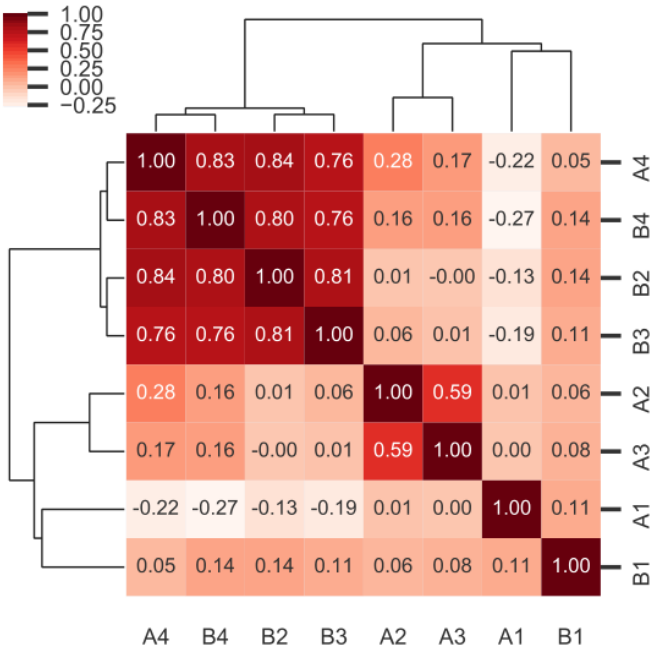


Figure 3.6 Pearson's correlation coefficient between the lengths of the eight active-site loops in 1,748 GH7s. The matrix of correlation coefficients is clustered so that loops with a similar pattern of correlation are grouped together. There is a high degree of positive correlation (darker red) between the lengths of the A4, B2, B3, and B4 loops.

One major advantage of the tree-based methods over other ML algorithms is the possibility of deriving and visualizing interpretable classification rules.[238, 239] In many applications of ML to biological problems, it is desirable to gain knowledge of biological relationships rather than merely apply ML as a predictive tool. Figure 3.5D shows rules derived from the single-node decision-tree classifiers trained on the A4, B2, B3, and B4 loops. A classification accuracy of 96.9% was achieved by the simple rule: if a GH7 has more than four residues in the B2 loop, then it is a CBH, else it is an EG. Overall, the

decision trees reveal that GH7 EGs tend to possess three or less residues in the B3 and B4 loops, four or less residues in the B2 loop, and five or less residues in the A4 loop.

Since the lengths of the A4, B2, B3, and B4 loops can independently discriminate between GH7 CBHs and EGs with accuracies greater than 94%, it is expected that there is a substantial degree of correlation between them. We conducted correlation analysis by computing the Pearson's correlation coefficient between the lengths of the eight loops of 1,748 GH7s (Figure 3.6). As expected, there is significant positive correlation between the lengths of the A4, B2, B3, and B4 loops ($r \geq +0.76$, $p < 0.0001$). The highest correlations are observed between the A4 and B2 loops (+0.84) and between the A4 and B4 loops (+0.83).

3.3.4 Discrimination of GH7 subtypes with position-specific classification rules: important residues for CBH/EG function

In discriminating GH7 CBHs and EGs with ML, we have used only the lengths of the active-site loops as features without considering the contributions of specific amino acids in the proteins. However, the interactions of specific residues are known to affect GH7 CBH/EG function, and mutagenesis studies have confirmed that certain positions play essential roles in GH7 activity.[41, 212, 213, 240] In this section, we investigate the relationships between specific residues in the proteins and the functional subtype.

It is common knowledge that although a protein's function arises from the combined effects of multi-level interactions between all residues in the protein, some residues contribute to function more significantly than others. Consequently, it is likely that in GH7s, if a position is considerably conserved in CBHs such that CBHs tend to

utilize a particular amino acid at that position, and EGs tend to not utilize the same amino acid at that position, or vice versa, then that position plays a vital role in the difference in CBH/EG function or structural stability. A typical example is position 40 (i.e. Trp40 in *Tre*Cel7A). From analysis of the structure-based MSA, we observe that this position is strongly conserved in CBHs with 92.5% exhibiting a Trp at this position, whereas it is notably variable in EGs with only 28.5% exhibiting a Trp at this position (Figure 3.7A). Considering only this clear difference in the amino acid distribution at position 40, we can infer that Trp40 likely contributes to CBH function. Mutation of Trp40 to Ala has, in fact, been shown to considerably decrease the activity of *Tre*Cel7A on crystalline cellulose but not on amorphous cellulose,[212] indicating that Trp40 is critical for processivity.[213] Consequently, we propose that applying a statistical method to mine for positions in GH7s that are conserved but have remarkably different amino acid distributions between CBHs and EGs can identify positions that play critical roles in CBH/EG function and processivity.

From the amino acid distribution at position 40, we obtain a single-node decision tree with the rule: Trp at position 40 implies CBH, else EG. This simple rule classifies 1,748 GH7 CBHs and EGs with an accuracy of 87.2%. Thus, a rational strategy for identifying positions likely associated with CBH/EG function is to derive similar rules for all positions in the MSA and select positions which yield high-performing rules. First, we split the MSA of 1,748 GH7 sequences into CBH and EG subalignments and then identified the consensus amino acid and the consensus amino acid type (i.e. aliphatic, aromatic, polar, positive, or negative) for each position in the subalignments. For each position, if X and Z are the consensus amino acids (or type) in the CBH and EG subalignment, respectively, we derived the following classification rules: X=>CBH and Z=>EG, X=>CBH and not

X=>EG, and not Z=>CBH and Z=>EG. Applying this strategy to 434 positions in the MSA (*Tre*Cel7A numbering), we derived 1,799 classification rules. For each rule, we measured the classification accuracy, sensitivity, specificity, and MCC, and tested the statistical significance by conducting chi-square test of independence. The 1,799 rules fairly have normally distributed MCC scores (Figure 3.7B), and the top five percent of rules (90 rules) have MCC scores of at least +0.73, and classification accuracies of at least 87% (Table 3.2 and Table A1.1 of Appendix A1, Figure 3.7 and Figure A1.2 of Appendix A1). These 90 rules are derived from 42 positions which are generally in close proximity to the cellodextrin ligand in the crystal structure. More than half of the top 90 rules are from positions within 5 Å of the cellononaose ligand bound in *Tre*Cel7A structure (PDB code: 4C4C). Moreover, most of the positions are closer to the tunnel entrance where cellulose chains are recruited by the enzyme for processive hydrolysis (Figure 3.7D and 3.7E).

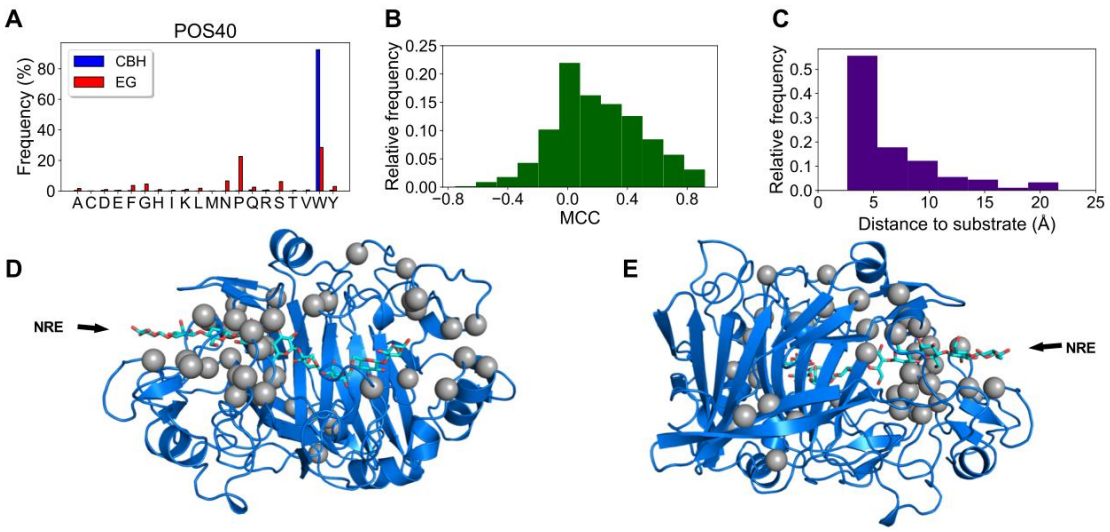


Figure 3.7 Top-performing position-specific classification rules for discriminating GH7 CBHs and EGs. (A) Amino acid distribution of GH7 CBHs and EGs at position 40 (*Tre*Cel7A numbering). Position 40 is strongly conserved as Trp in GH7 CBHs but not in

EGs. (B) MCC scores of 1,799 position-specific classification rules derived from the MSA. The top 90 rules have MCC scores of 0.73 or greater. (C) Histogram of minimum distance between the cellononaose ligand in *Tre*Cel7A (PDB code: 4C4C)[187] and positions from which the top 90 classification rules are derived. More than half of top 90 rules are derived from positions within 5 Å of the substrate. (D) Alpha carbons of 42 positions from which the top 90 classification rules are derived shown on the structure of *Tre*Cel7A. Most of these positions are near the substrate sites towards to the nonreducing end (NRE). (E) Posterior view of crystal structure.

Table 3.2 Top-performing position-specific classification rules relating amino acid residues and GH7 subtype (CBH/EG).All rules discriminate GH7 CBHs and EGs with accuracies of at least 87.0% and MCC scores of at least 0.73. Nearest distance to the nearest glycosyl residues was measured from the *Tre*Cel7A structure (PDB code: 4C4C). Statistical significance was tested by a chi-square test of independence. All rules are significant at $p<0.0001$. See Table A1.1 of Appendix A1 for rules between amino acid type and GH7 subtype. Positions from which the rules have been derived are shown on the crystal structure of *Tre*Cel7A in Figure 3.7.

*Tre*Cel7A position	Rule	Closest subsite	Distance to closest subsite (Å)	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
16	not Thr=>CBH, Thr=>EG	-2	19.0	97.7	82.6	93.9	0.83
37	Asn=>CBH, not Asn=>EG	-4	3.1	92.8	86.0	91.1	0.77
38	Trp=>CBH, not Trp=>EG	-4	3.2	93.2	96.2	93.9	0.85
39	Arg=>CBH, not Arg=>EG	-5	3.6	96.4	79.9	92.2	0.79
39	Arg=>CBH, His=>EG	-5	3.6	98.2	70.1	91.1	0.76
49	Asn=>CBH, not Asn=>EG	-7	2.7	90.9	85.3	89.5	0.73
51	Tyr=>CBH, not Tyr=>EG	-5	3.6	88.4	99.1	91.1	0.80
53	Gly=>CBH, not Gly=>EG	-5	4.9	90.6	90.7	90.6	0.77
56	Trp=>CBH, not Trp=>EG	-5	8.9	93.2	99.3	94.7	0.88
81	Thr=>CBH, not Thr=>EG	-5	4.1	88.1	91.2	88.9	0.74
82	Tyr=>CBH, not Tyr=>EG	-5	3.8	91.6	86.2	90.2	0.75
95	Phe=>CBH, not Phe=>EG	-4	7.0	84.0	97.5	87.4	0.74
97	Thr=>CBH, not Thr=>EG	-5	6.7	89.2	93.4	90.3	0.77
103	Asn=>CBH, not Asn=>EG	-5	2.7	92.1	87.8	91.0	0.77
105	Gly=>CBH, not Gly=>EG	-4	4.8	94.8	86.0	92.6	0.80
105	not Ser=>CBH, Ser=>EG	-4	4.8	99.7	74.9	93.4	0.82
105	Gly=>CBH, Ser=>EG	-4	4.8	97.2	80.4	93.0	0.81
106	Ser=>CBH, not Ser=>EG	-2	4.8	89.9	88.7	89.6	0.75
106	not Pro=>CBH, Pro=>EG	-2	4.8	99.2	86.9	96.1	0.89
106	Ser=>CBH, Pro=>EG	-2	4.8	94.5	87.8	92.8	0.81
120	Phe=>CBH, not Phe=>EG	-1	15.7	93.0	83.3	90.6	0.75
140	Leu=>CBH, not Leu=>EG	-1	8.5	83.2	98.2	87.0	0.73
146	Phe=>CBH, not Phe=>EG	-1	7.9	91.8	93.9	92.3	0.81
146	not Leu=>CBH, Leu=>EG	-1	7.9	94.3	79.4	90.6	0.75
146	Phe=>CBH, Leu=>EG	-1	7.9	93.1	86.7	91.4	0.78
179	Asp=>CBH, not Asp=>EG	-3	2.6	92.9	99.1	94.5	0.87
181	Lys=>CBH, not Lys=>EG	-5	2.8	92.0	99.3	93.8	0.86
192	Trp=>CBH, not Trp=>EG	-4	7.0	93.8	100.0	95.4	0.89
200	Asn=>CBH, not Asn=>EG	-4	3.5	85.5	99.3	89.0	0.77
202	Gly=>CBH, not Gly=>EG	-4	6.5	94.0	100.0	95.5	0.89
204	Gly=>CBH, not Gly=>EG	-4	10.7	95.0	99.5	96.2	0.91
251	Arg=>CBH, not Arg=>EG	2	3.4	86.9	99.8	90.2	0.79
262	Asp=>CBH, not Asp=>EG	2	4.1	95.1	97.3	95.7	0.89
262	not Gly=>CBH, Gly=>EG	2	4.1	98.7	69.0	91.2	0.76
262	Asp=>CBH, Gly=>EG	2	4.1	96.9	83.1	93.4	0.82
338	Phe=>CBH, not Phe=>EG	2	7.7	91.8	99.8	93.8	0.86
340	Asp=>CBH, not Asp=>EG	2	9.1	83.1	99.3	87.2	0.74
381	Tyr=>CBH, not Tyr=>EG	2	3.5	83.7	99.8	87.8	0.75
382	Pro=>CBH, not Pro=>EG	2	5.0	93.3	98.4	94.6	0.87
391	Gly=>CBH, not Gly=>EG	2	6.9	94.6	91.0	93.7	0.84
394	Arg=>CBH, not Arg=>EG	2	3.1	95.1	96.4	95.4	0.89
394	Arg=>CBH, Ala=>EG	2	3.1	97.4	72.6	91.1	0.76
396	not Pro=>CBH, Pro=>EG	2	12.9	85.5	96.6	88.3	0.75
401	not Glu=>CBH, Glu=>EG	-3	13.5	98.8	72.9	92.2	0.79
423	not Trp=>CBH, Trp=>EG	-1	18.1	97.4	72.6	91.1	0.76

3.3.5 Conserved aromatic residues in the active site of GH7s

GH7s possess several aromatic residues lining the active-site tunnel which have been suggested to play key roles in cellulolytic bond cleavage and processive action.[211] We have conducted bioinformatic analysis of conserved aromatic residues in the active site of GH7s. From the MSA of 1,748 GH7s, we selected positions that are located within 6 Å of the cellononaose substrate in the structure of *Tre*Cel7A (PDB code: 4C4C), and that have aromatic residues (Phe, Trp, Tyr, or His) at that position in the consensus sequence of CBHs or EGs (Figure A1.3 of Appendix A1). There are 17 of such aromatic positions in the MSA, and on the protein structure, these positions are distributed across the nine glycosyl subsites.

Furthermore, these 17 positions can be classified into three groups based on the conservation of aromatic amino acids (Table 3). The first group consists of positions that are conserved in both CBHs and EGs such that more than two-thirds of CBHs and EGs utilize aromatic residues at these positions. Positions 145, 171, 216, 228, 367, and 376 (*Tre*Cel7A numbering) fall in the first group. The second group consists of positions that are conserved as aromatic residues (>66%) in CBHs but not in EGs. Positions 38, 40, 51, 82, 252, 370, and 381 fall in the second group. The third group contains positions that are neither conserved (<66%) as aromatic residues in CBHs and EGs although the consensus amino acids are aromatic. Positions 39, 47, 53, and 247 fall in the third group.

When these positions are viewed on the crystal structure (Table 3.3, Figure 3.8), an interesting pattern is observed. Whereas positions that are strongly conserved in both CBHs and EGs (first group) are located near the catalytic center of the active site, positions that

are conserved in CBHs but not in EGs flank the catalytic center nearer to the "substrate-binding" sites (-7 to -1) or the "product-binding" sites (+1 to +2).

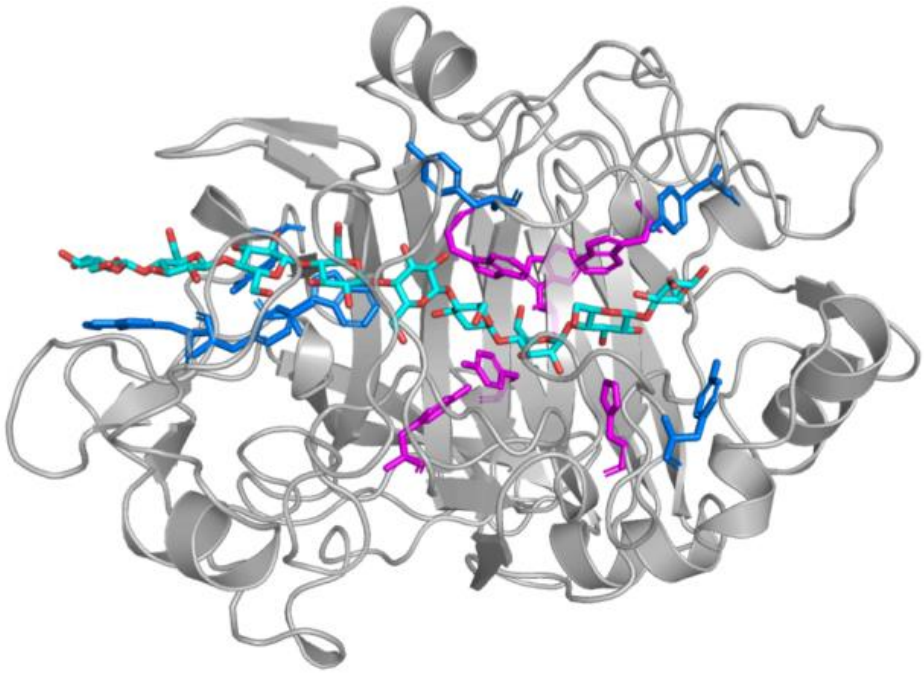


Figure 3.8 Conserved aromatic residues in the active site of TreCel7A (PDB code: 4C4C) within 6 Å of the cellononaose ligand. Residues in magenta are conserved (>66% frequency) in both GH7 CBHs and EGs and are found close to the catalytic center between -1 and +1 glycosyl subsites. Residues in blue are conserved in GH7 CBHs but not in EGs and flank the catalytic center.

Table 3.3 Positions within 6 Å of the cellononaose ligand in TreCel7A (PDB code: 4C4C) containing aromatic residues in consensus CBH or EG sequences. The positions are listed in order of proximity to the glycosyl subsites. Aromatic positions conserved in both CBHs and EGs are near the catalytic center, whereas aromatic positions conserved in only CBHs flank the catalytic center. All conserved positions are shown on the crystal structure of *Tre*Cel7A in Figure 3.8.

*Tre*Cel7A position	*Tre*Cel7A residue	CBH consensus residue	EG consensus residue	Frequency of aromatic residues in CBHs (%)	Frequency of aromatic residues in EGs (%)	Closest subsite	Distance to closest subsite (Å)	Aromatic residues conserved (>66%) in
47	S	Y	-	46.8	13.8	-7	3.9	None
40	W	W	W	93.3	36.0	-6	3.4	CBH
39	R	R	H	0.0	60.4	-5	3.6	None
53	G	G	W	0.3	29.4	-5	4.9	None
51	Y	Y	G	92.6	2.0	-5	3.6	CBH
82	Y	Y	Y	94.9	31.7	-5	3.8	CBH
38	W	W	A	94.2	29.4	-4	3.2	CBH
370	Y	H	E	87.3	1.8	-3	5.3	CBH
247	Y	Y	-	46.9	0.0	-2	2.7	None
145	Y	Y	Y	97.9	98.9	-2	2.7	CBH and EG
367	W	W	W	94.3	98.6	-1	3.0	CBH and EG
171	Y	Y	Y	98.0	97.3	-1	3.8	CBH and EG
216	W	W	W	97.9	68.1	-1	5.6	CBH and EG
228	H	H	H	96.6	97.5	1	2.8	CBH and EG
376	W	W	W	96.8	99.1	2	3.5	CBH and EG
252	Y	Y	-	85.0	17.2	2	5.8	CBH
381	Y	Y	-	92.8	0.2	2	3.5	CBH

3.3.6 Predicting the presence of CBMs with machine learning: relationships between the CD and the CBM

The CD of GH7 proteins may be attached to a second domain (the CBM) via a flexible linker. The CBM function is mostly attributed to enhancing the binding of the enzyme to the cellulose substrate, and thus, facilitating turnover by increasing enzyme concentration on the cellulose surface.[185]

We studied the distribution of family 1 CBMs in our dataset of 1,748 GH7s. First, a database of the 1,748 sequences was created and then a BLAST search of *Tre*Cel7A CBM was performed against the database. From a careful manual inspection of the BLAST alignment output, we selected an alignment score of 30 as the threshold so that GH7 sequences which yielded BLAST alignment scores of 30 or greater were determined to possess a family 1 CBM. We compared the distribution of CBMs among GH7 CBHs and EGs in our dataset and determined that 27% of GH7s contain a CBM, with 31% and 15% of GH7 CBHs and EGs exhibiting CBMs, respectively (Table 3.4). Thus, GH7 CBHs appear to be roughly two times more likely than EGs to contain a CBM. Moreover, a chi-square test of independence indicated that the relationship between CBM utilization and GH7 subtype (CBH/EG) is significant ($p < 0.001$).

To investigate relationships between the CD and the CBM, we applied ML to predict the presence of CBMs using the specific amino acid residues in the CD as features. Positions flanking the CD in the MSA were removed and one-hot encoding was applied to transform the amino acids in the MSA to binary variables.[241] Therefore, the MSA was transformed to a matrix such that the rows indicate the sequences, and columns denote the amino acid at positions in the MSA (features). Columns are labeled as “residue-position” and can take

values of 0 or 1. For example, a value of 1 at columns Q1 and S2 for *Tre*Cel7A indicate that Gln and Ser are present at positions 1 and 2 in the MSA, respectively (Figure 3.9B). Subsequently, one-hot encoding resulted in a high-dimensional matrix with 1,748 rows and 5,933 columns. We implemented the random forest algorithm[131] with 500 trees to predict the presence of a CBM using the 5,933 one-hot encoded features. The random forest algorithm is especially suitable for this classification problem because it is capable of robustly dealing with high dimensional data by performing implicit feature selection in the learning process,[135] is more tolerant to noise and overfitting,[131, 134] and can be used to evaluate the relative importance of the features.[136]

The performance of the random forest classifier was evaluated with 100 repetitions of five-fold cross validation with random undersampling, as described previously (Figure 3.4). Only 90% of the dataset was used for the cross validation; 10% of the dataset (174 sequences) was randomly selected and set aside for a separate final test. The random selection of the test dataset was implemented in such a way that a similar distribution (27% CBM, 73% no CBM) was maintained. In the validation routine, an accuracy of 90.8% was achieved by the 500-trees random forest trained on all 5,933 features (Table 3.6). A plot of the relative (Gini) importances[136] of the features shows that most of the 5,933 features contribute little or no information to the performance of the random forest classifier (Figure 3.9A). We reapplied the random forest algorithm using only the top 50 and the top 20 features with the highest Gini importances. The classifiers trained on only the top 20 and top 50 features showed fairly similar validation performance to the classifier trained on all 5,933 features (Table 3.6).

Table 3.4 Distribution of CBMs in GH7s showing the relationship between subtype (CBH/EG) and the presence of a CBM.GH7 CBHs are roughly two times more likely to possess a CBM than GH7 EGs (p<0.0001, chi-square test)

	CBH	EG	Total
Has CBM	407	66	473
No CBM	899	376	1275
Total	1306	442	1748
CBM frequency (%)	31.2	14.9	27.1

Table 3.5 Distribution of CBMs in GH7s showing the relationship between the presence of the rare disulfide bond (C4-C72 in TreCel7A) and the presence of a CBM. GH7s possessing this disulfide bond are roughly three times more likely to possess a CBM than GH7s lacking the disulfide bond (p<0.0001, chi-square test)

	Has disulfide bond	Lacks disulfide bond	Total
Has CBM	105	368	473
No CBM	54	1221	1275
Total	159	1589	1748
CBM frequency (%)	66.0	23.2	27.1

Table 3.6 Performance (%) of random forest classifiers in predicting presence of CBM.Validation and testing are performed on a 90%:10% split of the dataset, respectively. Validation performance is reported as mean ± standard deviation over 100 repetitions of five-fold cross validation.

	Validation				Testing
	All 5,933 features	Top 50 features	44 features (no C-terminus)	Top 20 features	Top 20 features
Accuracy	90.8 ± 2.1	90.9 ± 2.1	88.2 ± 2.5	89.3 ± 2.4	89.7
Sensitivity	93.7 ± 2.8	92.2 ± 2.9	89.6 ± 3.4	90.0 ± 3.2	95.7
Specificity	87.9 ± 3.5	89.7 ± 3.3	86.9 ± 3.7	88.5 ± 3.6	87.4
MCC	0.80 ± 0.05	0.81 ± 0.05	0.76 ± 0.05	0.78 ± 0.05	0.68

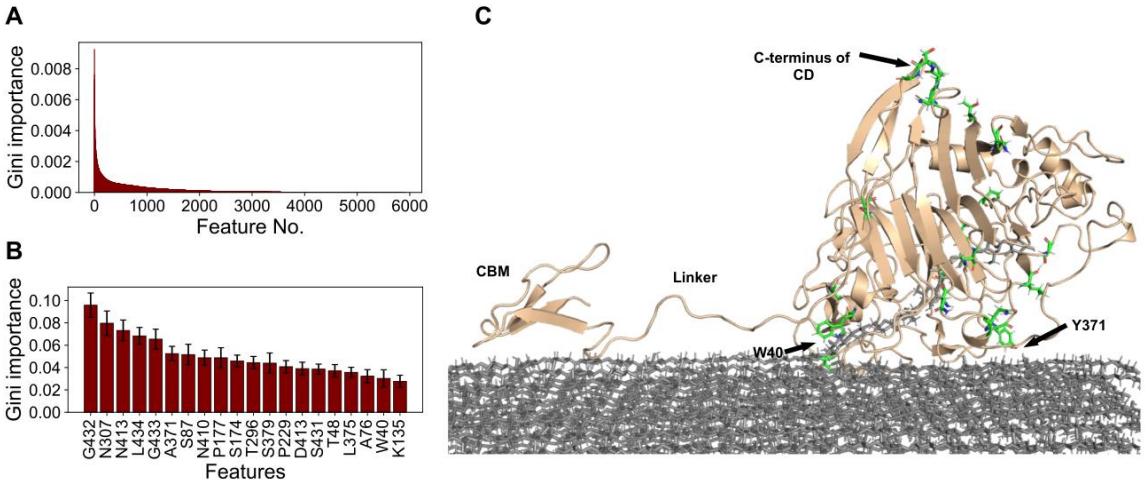


Figure 3.9 Top-performing features of the random forest classifier in predicting the presence of CBMs in GH7s. (A) Relative Gini importance of all 5,933 features derived from one-hot encoding of the MSA. Most features provide little information to the model (B) Relative (Gini) importance of top 20 features in the random forest classifier retrained on only top 20 features. Error bars indicate standard deviation measured over 100 repetitions of five-fold cross validation. (C) Residues of top 20 features (green sticks)

shown on the structure of *Tre*Cel7A (tan cartoon) on cellulose (gray sticks). The structure is derived from a snapshot (t = 0.73 μs) of MD simulations conducted in a previous work.[242]

Some residues at the C-terminus of the CD (where the CD connects with the CBM-linker domain) were identified to be among the most important positions in predicting the presence of CBMs (Figure 3.9B, Table A1.2 of Appendix A1). To confirm that the random forest algorithm was not predicting the presence of a CBM mainly by looking at these inter-domain connecting residues (S431, G432, S433, G433, T433, L434), we repeated the validation procedure with the top 50 features but excluded features derived from positions near the C-terminus (6 features removed, 44 features remaining). The results show that the performance of the new classifier trained on 44 features was only slightly lower, with the accuracy dropping by less than three percent. Moreover, on the separate test set, the classifier trained on the top 20 features achieved an accuracy of 89.7%, confirming that the presence of a CBM can be predicted from a few residues in the catalytic domain with considerable accuracy. In addition, we derived position-specific classification rules with each of the top 50 features, as described previously (i.e. X=>CBM, else, no CBM). As expected, all 50 rules independently performed worse, compared to the random forest classifier trained on all the 50 features (MCC <0.60, vs 0.81, see Table A1.2 of Appendix A1). Among these 50 rules, the top six rules are derived from L434, G433, T433, G432 (C-terminus residues), and C4 and C72, which are the Cys residues in *Tre*Cel7A that form a rare disulfide bridge (Table 3.5 and Figure A1.7 of Appendix A1).[46]

3.4 Discussion

In this study, we apply data mining techniques to investigate relationships between sequence and function of GH7s. We are able to accurately discriminate 1,748 GH7 CBHs and EGs with ML using only the number of residues in the active-site loops as features. However, whereas the ML models trained on the lengths of A4, B2, B3, and B4 loops achieved high predictive performance (>94% accuracy), the models trained on the other loops demonstrated mediocre or poor performance (Table 3.1, Figure 3.5A). These results indicate that the lengths of the A4, B2, B3, and B4 loops are primarily important for the difference in GH7 CBH and EG behavior. Greater exposure of the active site is generally accepted as a hallmark of nonprocessive cellulases (EGs). In addition, the ML results indicate that exposure of the active site in GH7 EGs occurs primarily at the product-binding region (+1 and +2 glycosyl subsites) due to deletions in the A4 and B4 loops, at the region below the catalytic center due to deletions in the B3 loop, and at the region to the lower left of the catalytic center due to deletions in the B2 loop (Figure 3.1 and 3.5C).

Earlier works have indicated that GH processivity correlates with ligand binding affinity, ligand solvation, and the flexibility of catalytic residues.[208, 243, 244] In *Tre*Cel7A, binding affinity is stronger at product-binding sites (+1, +2) than at the substrate-binding sites (-7 to -1), and this binding affinity difference has been proposed to be the driving force for the forward processive motion of the cellulase chain.[192, 208, 211, 245] Consequently, a logical explanation for why the lengths of the A4 and B4 loops strongly correlate with GH7 CBH/EG function is as follows: deletions in the A4 and B4 loops increase ligand solvation, disrupt protein-substrate hydrogen bonds, and lower binding affinity at the product binding sites, leading to a decrease in processivity. Similarly, the strong

relationship between the lengths of the B2 and B3 loops and GH7 CBH/EG function can be explained by the rationale that deletions in the B2 and B3 loops lead to an increase in solvation and a decrease in protein-ligand interactions in the substrate-binding sites, and an increase in solvation and flexibility of catalytic residues. It is interesting that although the A2 and A3 loops also overlay the catalytic center of the active site, their lengths show practically no correlation with GH7 CBH/EG function (Figure 3.5A and Figure A1.1 of Appendix A1), and exposure of the catalytic center in GH7s is achieved primarily by deletions in the B2 and B3 loops instead.

Moreover, the level of variation in lengths of the loops, as measured by the relative standard deviation, positively correlates with the predictive performance of the loops in discriminating GH7 CBHs and EGs (Figure 3.5A and 3.5B). This suggests that variation in the lengths of active-site loops was a major strategy in the evolutionary design of processivity in GH7s so that variation was allowed in the loops that significantly affect processivity and limited in other loops that have little impact on processivity (A2 and A3).

Furthermore, there is a strong positive correlation between the lengths of the A4, B2, B3, and B4 loops (Figure 3.6). Hence, in wild type GH7s, the shortening of any one of these four loops is highly associated with truncation of the other three loops. On our dataset of 1,748 sequences, we observed that if the B4 loop of a sequence is shortened, as is typical of GH7 EGs (i.e. possessing three residues or less), the probability that the A4, B2, and B3 loops are all shortened to typical GH7 EG lengths (i.e. five, four, and three residues or less, respectively) is 0.97 (Figure 3.5D). In other words, the pronounced concurrent shortening of the A4, B2, B3, and B4 loops observed in all crystal structures of GH7 EGs is remarkably conserved in EGs across the GH7 family.[188, 206, 246] This distinct bimodal distribution

(Figure A1.1 of Appendix A1) and strong correlation between loop lengths and GH7 subtype may serve as a valuable tool for correct gene annotation. Moreover, the strong conservation of loop lengths also indicate that there are coupled interactions between the A4, B2, B3, and B4 loops.[247] This might explain why in the recent work of Schiano-di-Cola *et al.*, independent deletions in the B3 and B4 loops did not lead to significant improvements in the activity of *Tre*Cel7A on amorphous cellulose, and deletions in the A4 loop rendered the enzyme inactive.[225] Cooperative synergy of deletions in other loops, as well as point mutations at key positions, may be required to fully exploit the effects of deletions in the B3, B4, and A4 loops.

Beyond the active-site loops, we have derived 90 position-specific classification rules from 42 positions in the MSA, such that the specific amino acid, or amino acid type, at any of these positions can independently predict the subtype of GH7s, with accuracies ranging from 87% to 97%. The high accuracy of these classification rules implies that there are strong constraints on the specific amino acids, or amino acid types, utilized by GH7 CBHs and EGs at these 42 positions. Such differential constraints likely signify that these positions play imperative roles in the difference between GH7 CBH and EG behavior. More than half of these positions are within 5 Å of the cellononaose substrate bound in the *Tre*Cel7A structure, and many of these positions cluster around the B2 loop (Figure 3.1, 3.9D and 3.9E). This finding provides a possible explanation for the observation that deletions in the B2 loop led to much greater changes in the CBH-behavior of *Tre*Cel7A than deletions in the B3 and B4 loops, relative to *Tre*Cel7B.[225] Since more of the important residues that yield high-accuracy position-specific classification rules cluster around the

B2 loop than other loops, deletions in the B2 loop likely leads to a disruption of a greater number of interactions necessary for CBH activity than deletions in the B3 and B4 loops.

Many of the 42 positions from which we derived classification rules have been identified and studied in previous works, and mutations at these positions have led to significant increase in catalytic efficiency.[43, 248, 249] Trp38, Tyr51, Asn103, Lys181, Asn200, Asp179, Arg251, Asp262, and Arg394 were identified in a docking study as residues that directly interact with and stabilize the cellulose substrate in the active site of *Tre*Cel7A.[250] Several of these positions have been further shown to form important stabilizing interactions with the substrate. Arg394 forms hydrogen bonds with the +2 glycosyl residue,[210, 211] Arg251 forms a salt bridge with Asp259 and hydrogen bonds with the +1 and +2 glycosyl residues,[41, 45, 211] and Asn103 and Lys181 form hydrogen bonds with the -5 glycosyl residue.[199, 251] Sørensen *et al*. studied mutants of *Rasamsonia emersonii* Cel7A in which two Asn residues on the B2 loop, Asn194 and Asn197 (Asn197 and Asn200 in *Tre*Cel7A, respectively) were replaced with Ala.[43] They observed that the mutations led to a decrease in substrate affinity and processivity, thus, enabling faster enzyme-substrate dissociation and a corresponding increase in activity on crystalline cellulose. In this present work, the Asn200 position yields the following classification rule: Asn implies CBH, and not Asn implies EG, which discriminates GH7 CBHs and EGs with an accuracy of 89%. Similarly, Bu *et al.* conducted computational studies of several *Tre*Cel7A residues including Arg251, Asp262, and Tyr381.[45] These residues were identified to substantially interact with the cellobiose substrate and mutation to Ala resulted in considerably weaker binding of cellobiose in the product-binding site. It was suggested that these mutants would demonstrate improved biomass conversion efficiency due to

accelerated expulsion of the cellobiose product. In this present study, these positions (251, 262, and 381) also yield high-accuracy classification rules with accuracies of at least 88%. Additionally, Mitsuzawa *et al.* determined that mutation of Asn63 and Lys203 to Ala in *Talaromyces cellulolyticus* Cel7A (Asn37 and Lys181 in TreCel7A, with classification accuracies of 91% and 94%, respectively) led to a remarkable increase in activity on cellulose.[248]

Some positions farther away from the active site also yielded high-accuracy classification rules. For example, position 401 – conserved as Ser in CBHs but as Glu in EGs and more than 13 Å away from the cellodextrin ligand in the *Tre*Cel7A structure – generates a CBH/EG classification rule with an accuracy of 92%. Although residues at positions such as 401 may not directly interact with the cellulose substrate in the active site, they may participate in long-range interactions that affect GH7 CBH and EG behavior. Further studies are required to determine the specific roles these conserved positions play in function and structural stability. Altogether, we surmise that the positions that yield high-accuracy classification rules play key roles in GH7 CBH/EG function and, as such, should be carefully considered when engineering the protein at or around these sites.

Bioinformatic analysis of the MSA revealed conserved aromatic positions in the active site that are within 6 Å of the cellulose substrate in *Tre*Cel7A (Table 3.3, Figure 3.8). The results indicate that whereas conserved aromatic residues in the active site of GH7 CBHs span the entire active-site tunnel, conserved aromatic residues in the active site of GH7 EGs are clustered around the catalytic center. Moreover, aromatic positions near the catalytic center are conserved in both GH7 CBHs and EGs. This arrangement of conserved aromatic residues in the active site suggests that while aromatic residues near

the catalytic center (Y145, W216, H228, W367, and W376) play major roles in catalytic bond cleavage, conserved aromatic residues that flank the catalytic center (W38, W40, Y51, Y82, Y252, Y370, and Y381) are utilized mainly by CBHs for processive motion. Several experimental and computational studies support this hypothesis.[45, 213, 252, 253]

Taylor *et al*. assayed chimeras derived from interchanged subdomains of *Pfu*Cel7A and *Tre*Cel7A. Although the CD of *Pfu*Cel7A exhibited greater efficiency on biomass than the CD of *Tre*Cel7A, interchanging CBM and linker regions did not yield a uniform trend in catalytic efficiency. As a result, it was concluded that there are complex interactions that are not yet well-understood between the domains.[46] In this work, we have applied ML to predict the presence of CBMs from amino acid positions in the CD to map relationships between the CBM and CD of GH7s. First, our data indicate that GH7 CBHs are roughly two times more likely to utilize CBMs than GH7 EGs, which is as expected since CBMs likely enable CBHs stay longer on the cellulose substrate to facilitate consecutive hydrolysis. Furthermore, ML results show that the presence of a CBM in GH7s can be accurately predicted (89.3%) using only 20 features derived from 19 positions in the catalytic domain (Table 3.6). This high predictive accuracy largely suggests that there are constraints and key functional relationships between residue positions in the CD and the presence of a CBM in the gene. Interestingly, on the protein structure, these 19 positions are mostly located on loops or at turns all over the protein structure (Figure 3.9C). Moreover, the amino acid residues constituting the 20 features are mostly small amino acids (such as Gly, Ser, Thr, Asp, and Asn) that are known to affect the conformational flexibility of proteins.[254, 255] Taken together, our ML results, while preliminary, suggest that the presence of CBMs in GH7s correlates with the overall conformational flexibility

of the CD, and that CBMs may exist, in part, to compensate for highly flexible CDs that are more likely to detach from the cellulose surface. Moreover, the position-specific rules we derived from the top 50 random-forest features in predicting CBMs indicate that GH7s possessing a rare disulfide bond (C4-C72 in *Tre*Cel7A) are about three times more likely to possess a CBM than GH7s that lack this disulfide (Table 3.6, Table A1.2 and Figure A1.7 of Appendix A1). In a previous work, mutation of C4 and C72 in *Tre*Cel7A was shown to increase cellulolytic efficiency and flexibility of the tunnel entrance.[46] Since an extra disulfide bridge would generally decrease the flexibility of the CD, the correlation of C4 and C72 with the presence of a CBM is contrary to our hypothesis that CBMs compensate for the flexibility of the CD. This paradoxical correlation, thus, warrants further experiments to investigate such relationships.

In conclusion, we have used ML to uncover key positions in GH7 sequences that appear to be related to function and statistical relationships between GH7 sequence and functional diversity. While these relationships are statistically significant, we stress that they may be influenced by sampling and phylogenetic biases inherent to the dataset. Nonetheless, as the ML strategies we have applied to GH7s may be extended to other protein families, particularly where multiple functional classes exist in the family (such as CBH/EG or CBM/no-CBM), this work provides a solid basis for the statistical investigation of sequence-function relationships in protein families. We also anticipate that the findings in this work will inform further propitious studies for the design of more efficient cellulases.

3.5 Materials and methods

3.5.1 Sequence datasets

Sequences were retrieved by protein-protein BLAST searches against the NCBI non-redundant database by using *Tre*Cel7A (P62694.1) and *Fox*Cel7B (AAA65586.1) as query sequences. BLAST search was implemented with the NCBI web server (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using default settings. Only sequences with E-values of 1e-20 or better and query cover of 60% or more were retained. The query cover threshold of 60% was applied to exclude the large number of fragment sequences returned by the BLAST search. A total of 2,024 sequences were retrieved. A sequence identity threshold of 99% was applied to remove redundant sequences so that only 1,748 sequences were left in the dataset. From manual inspection of the BLAST output, 60 of these sequences consisted of multiple domains other than GH7. Other domains were deleted in these sequences leaving only one GH7 domain for each sequence. The SwissProt/UniProt dataset of 44 sequences was obtained by a similar BLAST search against the SwissProt/UniProt database. All datasets and Python scripts used in this study are available at https://github.com/jafetgado/Cel7ML.

3.5.2 Sequence alignments

Sequence alignments of the SwissProt/UniProt dataset (44 sequences) and the annotated NCBI dataset (427 sequences) were conducted with MAFFT version 7[256] using BioPython[169] with default settings. Due to the greater diversity of the larger dataset (1,748 sequences), in order to avoid generating erroneous alignments, a structure-based sequence alignment was implemented for the larger dataset. First, structural alignment of 20 GH7

structures (16 CBHs, 4 EGs) was conducted with the Promals3D web server.[257] The structural alignment was manually edited in UGENE[258] following standard manual adjustment methods.[259] Then, an MSA of the 1,748 sequences was generated with the MAFFT add-sequences option[256] by adding the sequences to the structural alignment. Sequence alignments were viewed with ESPript (http://espript.ibcp.fr),[260] and sequence logos (Figure A1.4 of Appendix A1) were generated with WebLogo (https://weblogo.berkeley.edu/logo.cgi).[261]

3.5.3 Machine learning and performance evaluation

Profile hidden Markov models were constructed from the MSAs with a local version of the HMMER software (version 3.1b2).[229, 262] All ML methods were implemented using the Scikit-learn Python package (version 0.20.3).[108] The K-nearest neighbor (KNN) classifier was trained with the "n_estimators" parameter (k) set to an optimal value of 10 (best of 5, 10, and 15). A radial basis function (RBF) kernel was applied in the support vector machine (SVM) classifiers, and default settings were used for the logistic regression classifiers. To avoid overfitting with the decision trees, the depth of the trees was limited to the number of features. Hence, single-feature decision trees had a "max-depth" of one, and the decision tree trained on all eight features had a "max-depth" of eight.

There were severe outliers in the lengths of active-site loops that would have skewed the ML results. For example, from the MSA, a sequence (GenBank accession: CRK24563.1) had 140 residues in the B2 loop. These extremities may have resulted from sequencing or splicing errors. Before the ML procedure, outliers were capped to an

arbitrarily selected maximum limit (Figure A1.5 of Appendix A1).[106] All non-binary features applied in ML were standardized by converting them to Z-scores according to equation 2.3.

The ML algorithms were applied to discriminate between a positive class (CBH or CBM) and a negative class (EG or no CBM), resulting in four classification outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The performance of the ML algorithms was evaluated by computing the sensitivity, specificity, accuracy, and MCC according to the equations 2.7, 2.8, 2.6, and 2.13, respectively.

CHAPTER 4. Enabling Microbial Syringol Conversion through Structure-guided Protein Engineering.

This chapter has been reprinted with permission from Machovina et al., Copyright 2019,[55] National Academy of Sciences USA. The findings presented in this chapter are a large collaborative work comprising computational, structural, and experimental studies of a cytochrome P450 demethylase (GcoA) for conversion of lignin subunits enabled the engineering of GcoA for expanded specificity and activity. The author of this dissertation performed the conservation analyses, which provided helpful insight for mutagenesis and other parts of the study. Expression and biochemical characterization were performed by collaborators (Melodie M. Machovina, April Oliver, Christopher Johnson) at Montana State University and the National Renewable Energy Laboratory (NREL). Crystallization and structural studies were performed by collaborators at University of Portsmouth (Sam J.B. Mallinson and Daniel J. Hinchen). Other computational calculations were performed by collaborators at NREL and University of California, Los Angeles (Brandon C. Knott, Lintao Bu, Michael F. Crowley, Alexander W. Meyers, Graham Schmidt, Marc Garcia-Borras, and Kendall N. Houk).

4.1 Abstract

Microbial conversion of aromatic compounds is an emerging and promising strategy for valorization of the plant biopolymer lignin. A critical and often rate-limiting reaction in aromatic catabolism is *O*-aryl-demethylation of the abundant aromatic methoxy groups in lignin to form diols, which enables subsequent oxidative aromatic ring-opening.

Recently, a cytochrome P450 system, GcoAB, was discovered to demethylate guaiacol (2-methoxyphenol), which can be produced from coniferyl alcohol-derived lignin, to form catechol. However, the native GcoAB has minimal ability to demethylate syringol (2,6-dimethoxyphenol), the analogous compound which can be produced from sinapyl alcohol-derived lignin. Despite the abundance of sinapyl alcohol-based lignin in plants, no pathway for syringol catabolism has been reported. Here, we employed structure-guided protein engineering to enable microbial syringol utilization with GcoAB. Specifically, a phenylalanine residue (GcoA-F169) interferes with the binding of syringol in the active site, and upon mutation to smaller amino acids, efficient syringol *O*-demethylation is achieved. Crystallography indicates that syringol adopts a productive binding pose in the variant, which molecular dynamics simulations trace to the elimination of steric clash between the highly flexible side chain of GcoA-F169 and the additional methoxy group of syringol. Lastly, we demonstrate *in vivo* syringol turnover in *Pseudomonas putida* KT2440 with the GcoA-F169A variant. Taken together, this study highlights the significant potential and plasticity of cytochrome P450 aromatic *O*-demethylases in the biological conversion of lignin-derived aromatic compounds.

4.2 Significance

Lignin is an abundant but underutilized heterogeneous polymer found in terrestrial plants. In current lignocellulosic biorefinery paradigms, lignin is primarily slated for incineration, but for a non-food plant-based bioeconomy to be successful, lignin valorization is critical. An emerging concept to valorize lignin employs aromatic-catabolic pathways and microbes to funnel heterogeneous lignin-derived aromatic compounds to

single high-value products. For this approach to be viable, the discovery and engineering of enzymes to conduct key reactions is critical. In this work, we have engineered a two-component cytochrome P450 enzyme system to conduct one of the most important reactions in biological lignin conversion, namely aromatic *O*-demethylation of syringol, the base aromatic unit of S-lignin, which is highly abundant in hardwoods and grasses.

4.3 Introduction

Lignin is a heterogeneous, recalcitrant biopolymer that is prevalent in plant cell walls, where it provides structure, defense against pathogens, and water and nutrient transport through plant tissue.[263] Lignin is synthesized primarily from three aromatic building blocks,[56, 263] making it the only abundant and renewable aromatic carbon feedstock available. Due to its recalcitrance, rot fungi and some bacteria have evolved powerful, oxidative enzymes that deconstruct lignin to smaller fragments.[53, 54] Once broken down, the lignin oligomers can be assimilated as a carbon and energy source through at least four known aromatic-catabolic pathways.[56, 57]

A critical reaction in the aerobic catabolism of lignin-derived compounds is *O*-aryl-demethylation, which occurs on methoxylated lignin-derived compounds to produce aromatic diols such as catechol (1,2-dihydroxybenzene), protocatechuate (3,4-dihydroxybenzoate), and gallate (3,4,5-trihydroxybenzoate). Next, the aromatic rings are cleaved by intradiol or extradiol dioxygenases, and the products are funneled into central metabolism.[51, 264] Harnessing this catabolic capability for transforming heterogeneous lignin streams into valuable chemicals is of keen interest,[51, 265-269] and essential for economical lignocellulose conversion [268, 270, 271].

In most plants, lignin comprises primarily coniferyl (G) and sinapyl (S) alcohol monomers, which have either one or two methoxy groups on the aryl ring, respectively. Nearly all lignin-derived aromatics require *O*-demethylation of these methoxy groups as an essential step in their conversion to central intermediates. Significant effort has been dedicated to the discovery of enzymes that can demethylate the methoxy substituents of diverse aromatic compounds.[68, 69, 272-281] Ornston *et al.* characterized the *O*-demethylation of vanillin (4-hydroxy-3-methoxybenzaldehyde) and vanillate (4-hydroxy-3-methoxybenzoate) analogs by the VanAB monooxygenase from *Acinetobacter baylyi* ADP1, which contains a Rieske nonheme iron center [274, 275]. The three-component LigX monooxygenase system from *Sphingobium* sp. SYK-6, described by Masai *et al*., also contains a Rieske nonheme iron component,[277] that is responsible for *O*-aryl-demethylation of a model biphenyl compound that mimics those in lignin. Masai *et al.* additionally described two tetrahydrofolate-dependent enzymes, LigM and DesA, responsible for *O*-aryl-demethylation of vanillate and syringate (4-hydroxy-3,5-dimethoxybenzoate), respectively.[68, 276] Cytochrome P450 systems have also been reported to demethylate aromatic compounds such as guaiacol, 4-methoxybenzoate, and guaethol (2-ethoxyphenol).[69, 273, 280, 281] However, the full gene sequences were either unreported or only recently identified,[69, 273, 282] or the substrate was not of direct interest to lignin conversion.[280, 281]

Our recent characterization of a two-component P450 enzyme system, consisting of a reductase, GcoB, and P450 oxidase, GcoA, demonstrated that it demethylates diverse aromatic compounds including guaiacol (which can be derived from coniferyl alcohol and represents the aromatic functionality of G-lignin that must undergo demethylation),

guaethol, anisole (methoxybenzene), 2-methylanisole, and 3-methoxycatechol (3MC),[70] with similar or greater efficiency than other *O*-aryl-demethylases described in the literature.[277, 279, 283] However, GcoAB showed poor reactivity towards syringol, which can be derived from sinapyl alcohol via high-temperature reactions and represents the aromatic functionality of S-lignin that must undergo *O*-demethylation for further catabolism via ring-opening dioxygenases. Together, G- and S-lignin are the major components of lignin in hardwoods and grasses.[263] Due to their abundance, it is important to find enzymes that can act on the methoxy groups of both G- and S-lignin subunits. To date, there are no reports describing syringol *O*-demethylation or more broadly, even its catabolism by microbes. Rather, the best-studied biological reaction of syringol is its 4-4 dimerization to form cerulignone.[284-287]

Though our prior work showed that GcoA was not effective for syringol *O*-demethylation, crystallographic studies and molecular dynamics (MD) simulations indicated that a triad of active site phenylalanine residues is both highly mobile and important for positioning the substrate in its catalytically competent pose. In this study, we hypothesized that substitution of GcoA-F169, which has the closest interaction with the bound substrate, may relax the specificity of the enzyme sufficiently to permit the *O*-demethylation of S-lignin type substrates. We tested that hypothesis using biochemical, structural, computational, and *in vivo* approaches. We demonstrate highly efficient *in vitro* and *in vivo* syringol turnover through structure-guided protein engineering, where the enzyme also retains highly efficient activity toward guaiacol.

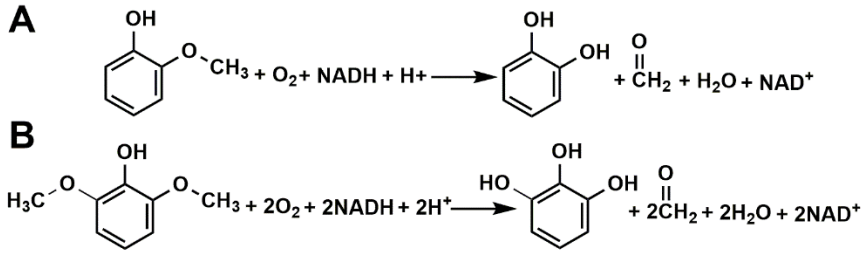
4.4 Results

4.4.1 The syringol binding mode can be modulated by active site engineering.

Guaiacol assumes a productive orientation in the active site of GcoA, resulting in a shift in the spin state of the heme iron from low (S=1/2) to high (S=5/2), due to the action of amino acid side chains that create a tight-fitting hydrophobic pocket. The closest contact is with GcoA-F169, which forms a hydrophobic interaction with the C6 carbon on the aromatic ring of guaiacol. Prior MD simulations suggested that this residue is highly mobile, predicting that the productive complex forms dynamically.[70] Superposition of the co-crystal structures of GcoA with guaiacol and syringol reveals a shift in the positions of GcoA-F169 and the reactive syringol methoxy group relative to the heme (see below). Functionally, these shifts in the GcoA-F169 position permit binding of syringol, though binding in the productive conformation as measured by the shift in Fe(III) spin state is substantially diminished relative to binding of guaiacol (Table 4.1).[70] We hypothesized that mutation of GcoA-F169 to a smaller residue (alanine) may relieve the apparent steric clash between it and the bound ligand in the active site, allowing syringol to adopt a productive conformation. The GcoA-F169A variant was therefore prepared and its guaiacol and syringol binding properties measured (Figure A2.1 of Appendix A2). Both ligands were able to stimulate the Fe(III) spin state conversion at levels close to wild-type (WT) (Table 4.1), with K_D values in the low micromolar range. Notably, 3MC also bound and induced a spin state change in Fe(III), though with less affinity than guaiacol or syringol. We concluded that active site engineering could indeed lead to productive syringol binding and potentially turnover by GcoA-F169 variants, and therefore subsequently studied their reactivity with guaiacol, syringol and 3MC.

4.4.2 GcoA-F169A efficiently demethylates both guaiacol and syringol with only limited uncoupling.

Substrate analogs are known to stimulate the P450 reaction with NADH/O_2 without concomitant oxygenation of the organic substrate. This leads to uncoupling of the NADH and substrate oxidation reactions, and reduction of O_2 either to H_2O_2 or to H_2O (34). Prior work showed that syringol led to stimulation of NADH consumption by WT GcoA (28), though without substantial syringol turnover. To address whether guaiacol and/or syringol would serve as substrates of GcoA-F169A, we monitored the disappearance of NADH (UV/vis) and aromatic substrate (HPLC) over time (Scheme 4.1). The rates of organic substrate and NADH consumption were robust and the same within error, regardless of whether guaiacol or syringol was used (Table A2.1 of Appendix A2). This suggests that both guaiacol and syringol serve as substrates for GcoA-F169A.



Scheme 4.1 *O*-demethylation of (A) guaiacol to form catechol and formaldehyde or (B) syringol to form pyrogallol and two formaldehydes.The singly demethylated species, 3-methoxycatechol (3MC), is expected to form as an intermediate in reaction B). See Figure 4.1.

Table 4.1 Efficacy of GcoA-F169A relative to WT GcoA in binding and demethylating guaiacol and syringol.

	WT GcoA			GcoA-F169A		
	guaiacol	syringol	3MC	guaiacol	syringol	3MC
K_D (μM)[a]	0.0065 ± 0.002	2.8 ± 0.5	3.7 ± 0.1	7.1 ± 0.1	1.7 ± 0.07	9.5 ± 0.2
%Fe(III) spin state conversion[a]	87 ± 1	56 ± 2	62 ± 0.01	72 ± 0.3	76 ± 0.7	65 ± 0.6
k_{cat} (sec^{-1})[b]	6.8 ± 0.02	-	n/a[d]	11 ± 0.03	5.9 ± 0.01	n/a
K_M (mM)[b]	60 ± 10	-	n/a	40 ± 6	10 ± 1	n/a
k_{cat}/K_M ($mM^{-1}sec^{-1}$)[b]	110 ± 20	-	n/a	290 ± 40	600 ± 90	n/a

[a]K_D was measured by titrating in 0-60 μM of substrate into a solution containing 2-6 μM WT or F169A GcoA in air saturated buffer (25 mM HEPES, 50 mM NaCl, pH 7.5, 25 °C) and recording the ferric spin state change from the low spin (417 nm) to high spin (388 nm) species. The % spin state conversion was calculated by dividing the final high spin species (max absorbance at 388 nm) by the starting low spin species (max absorbance at 417 nm).

[b]The Michaelis constants are apparent as the dioxygen and GcoB concentrations are not known to be saturating. The conditions used were: 0.2 μM GcoAB, 100 μg/mL catalase, 300 μM NADH, 210 μM O_2, and 0-300 μM substrate, 25 °C, 25 mM HEPES, 50 mM NaCl, pH 7.5.

[C]n/a and dashes: Michaelis constants for 3MC could not be directly measured because of a high level of NADH uncoupling, indicated by "n/a". See text. A substantial amount of syringol turnover was not observed for WT GcoA, again making it impossible to measured Michaelis parameters. This is indicated by dashes.

The oxidative *O*-demethylation of guaiacol moreover appeared to be largely coupled to NADH consumption. When NADH and O_2 were present in excess of guaiacol, the measured stoichiometry of the GcoA-F169A-catalyzed reaction was very close to one molecule each of guaiacol and NADH consumed to one formaldehyde and one catechol produced (Figure 4.1), without overconsumption of NADH (103 ± 7% coupling efficiency, Table A2.1 of Appendix A2).

Since both methoxy groups of syringol can potentially serve as substrates, we examined syringol turnover in a number of ways. Syringol (100 μM) was first incubated with NADH (200 μM) and excess dissolved O_2 (210 μM), and the reaction with the GcoA-F169A mutant allowed to go to completion. As with the guaiacol reaction, all of the NADH and syringol were consumed (Figure 4.1), implying that syringol undergoes two *O*-demethylations, producing 3MC and then pyrogallol. However, less formaldehyde (170 ± 10 μM) was produced than expected, suggesting some uncoupling of NADH/O_2 consumption from the oxidative *O*-demethylation. Consistent with that hypothesis, 50 ± 4 μM of the singly demethylated intermediate 3MC was observed at the end of the reaction, even though sufficient NADH/O_2 were present to enable its complete conversion to pyrogallol. Notably, pyrogallol was not detected under any of the conditions used here, possibly due to its well-known instability in the presence of O_2 (35).

The stoichiometric analysis was next repeated with NADH and syringol present in equal concentrations (~200 μM each; 210 μM O_2), conditions expected to permit at most half of the available methoxy groups to react. All of the NADH and 150 ± 6 μM of syringol were consumed; 200 ± 3 μM formaldehyde and 120 ± 2 μM 3MC were generated (Figure 5.1, Table A2.2 of Appendix A2). The accumulation of roughly half an equivalent of 3MC

(relative to NADH) under these conditions suggested that the first *O*-demethylation of syringol must be faster than the second, and that the uncoupling reaction was likely stimulated by 3MC rather than by syringol. Consistent with those expectations, the rate of 3MC disappearance measured by HPLC was significantly slower than the disappearance of either guaiacol or syringol (2.6 ± 0.3 μM 3MC s^{-1} μmol GcoA-F169A^{-1} *versus* 5.1 ± 0.8 μM syringol s^{-1} μmol GcoA-F169A^{-1}, Table A2.1 of Appendix A2); moreover, the faster consumption of NADH relative to 3MC suggested diminished reaction coupling (64 ± 10% coupling, Table A2.1 and A2.2 of Appendix A2). In reactions containing 100 μM of 3MC and 200 μM NADH, the majority of the initially available NADH was consumed, and approximately 100 μM of formaldehyde was produced (Figure 4.1). We hypothesized that the observed overconsumption of NADH was due to the uncoupled reaction, leading to H_2O_2 production. The production of H_2O_2 in the presence of excess NADH/O_2 and limiting 3MC was confirmed using Amplex Red and horseradish peroxidase (Table A2.3 of Appendix A2). As a consequence, accurate values for k_{cat} and K_M could not be measured using 3MC as a substrate (Table 4.1).

A broader survey of variants at the GcoA-F169 position (amino acids S, H, V, I, and L in addition to A) confirmed that GcoA-F169A exhibits the best catalytic performance both in terms of specific activity and reaction coupling, although other small side chains (S, V) also permitted reactivity with syringol, suggesting these may permit syringol to assume a reactive conformation at the heme. Apparent steady state kinetic parameters measured in air and at potentially sub-saturating concentrations of GcoB (Table 4.1, Figure A2.2 and Table A2.2 of Appendix A2) suggest that GcoA-F169A is a more effective catalyst toward the first methoxy group of syringol relative to guaiacol, with

k_{cat}/K_M[syringol] nearly double k_{cat}/K_M[guaiacol]. Moreover, GcoA-F169A has a slightly improved performance with guaiacol as a substrate relative to the WT enzyme.

4.4.3 Structural analysis reveals productive syringol reorientation in GcoA-F169 variants.

Superposition of the structures of GcoA-ligand complexes indicated a significant rotation and translation of bound syringol relative to guaiacol (Figure 4.2A), and we hypothesized that this could form the basis for the unproductive syringol uncoupling in the native enzyme. The comparative distances between the heme and the proximal methoxy carbon of guaiacol vs syringol are within 0.4 Å, and even closer between the heme and methoxy oxygens (within 0.1 Å). In addition, there is no significant deviation from the plane of the aromatic rings between these ligands. In contrast, the angle of presentation of the methoxy to the heme diverges significantly in these complexes. Using the angle between the methoxy oxygen, heme iron, and terminal methoxy carbon atoms (O-Fe(III)-C) as a convenient readout of relative orientation, there is a 55% increase in angle from the guaiacol-bound structure (8.3˚) compared with the syringol-bound structure (12.9˚).

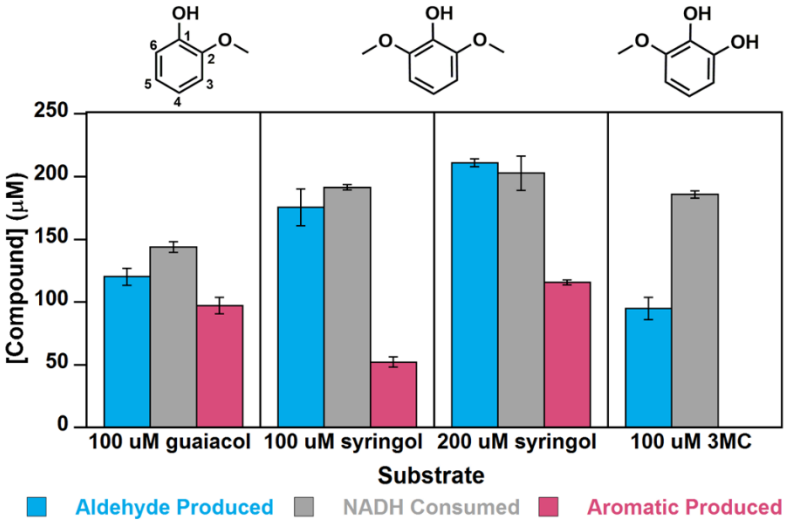


Figure 4.1 Quantitative analyses of substrate consumption and product generation indicate nearly complete coupling of NADH/O2 consumption to substrate O-demethylation for guaiacol, and progressively more uncoupling for syringol and 3MC. NADH (200 µM) and guaiacol, syringol, or 3MC (100 or 200 µM) were incubated in air with 0.2 µM GcoA-F169A and GcoB (each) (25 mM HEPES, 50 mM NaCl, pH 7.5, 25°C, 210 µM O_2). Reactants and products were quantified when the UV/vis spectrum ceased changing and the reaction was deemed complete. The total NADH consumed is compared above to the amounts of formaldehyde and de-methylated aromatic compound produced. Pyrogallol, the *O*-demethylated product of 3MC,is unstable in air under the conditions used in the assay and was not detected. Error bars represent ± 1 standard deviation from three or more independent measurements. P-values comparing NADH consumption and formaldehyde production were 0.035, 0.20, 0.41, and 0.0035 for guaiacol, 100 µM syringol, 200 µM syringol, and 3MC, respectively. For NADH consumption and aromatic product production, these values were 0.011, 0.00031, and 0.0084 for guaiacol, 100 µM syringol, 200 µM syringol, respectively.

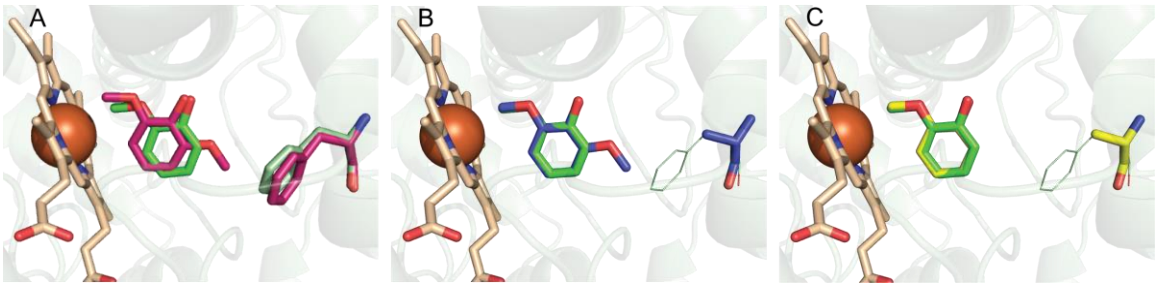


Figure 4.2 Superpositions of WT and GcoA-F169A ligand-bound structures of GcoA, the P450 monooxygenase component of GcoAB are shown. The heme is colored in bronze stick. (A) The guaiacol (green, PDB ID 5NCB) and syringol (pink, PDB ID 5OMU) complexes with WT GcoA are shown with the position of the GcoA-F169 residue highlighted. The translation and rotation of syringol compared to guaiacol results in a shift of the target methoxy carbon away from the heme; Fe(III) to guaiacol methoxy carbon distance is 3.9 Å; Fe(III) to proximal syringol methoxy carbon distance is 4.3 Å . (B) The engineered GcoA-F169A-syringol complex (blue) enables positioning of the reactive methoxy group relative to the heme in a mode consistent with productive guaiacol binding. GcoA-F169 from the guaiacol-bound WT structure is shown in green lines. (C) Superposition of the WT (green) and GcoA-F169A (yellow) guaiacol-bound complexes reveals that guaiacol sits in an identical position in both crystal structures. GcoA-F169 from the guaiacol bound WT structure is shown in green lines.

To investigate this further, we generated multiple high-resolution GcoA-F169 variant co-crystal structures. A set of syringol-bound structures (Table A2.4 and Figure A2.4 of Appendix A2) provided direct insight into the minimal reduction in side-chain bulk required to achieve the productive binding mode equivalent to that of guaiacol. A step-wise trajectory of the bound syringol towards this optimum orientation with decreasing side-

chain bulk was observed in the superposition of four co-crystal structures (Figure A2.3A of Appendix A2). Specifically, GcoA-F169H creates an improved substrate orientation, further improved by GcoA-F169V, and essentially optimized in both the GcoA-F169S and F169A proteins (Figure A2.3A of Appendix A2). Indeed, a comparison of the GcoA-F169A-syringol structure with the WT guaiacol structure revealed an almost perfect alignment relative to the aromatic rings of each substrate and a methoxy O-Fe-C angle of 8.6˚, within 0.3° of that observed for guaiacol (Figure 4.2B).

Each protein variant also crystallized successfully with guaiacol (Table A2.5 and Figure A2.5 of Appendix A2) and the structures showed that the orientation of the bound ligand remained consistent with that of the WT enzyme (Figure A2.3B of Appendix A2). Even the largest reduction in side-chain bulk, represented by the GcoA-F169A variant, retained the ideal reactive geometry for the natural substrate (Figure 4.1C). Furthermore, comparison of the surrounding active site architecture confirmed no significant deviation from the WT. The resolution of these structures (1.66-2.17 Å) also provides sufficient electron density quality to explore changes in the hydration of the pocket (Figure A2.6 of Appendix A2). While the native enzyme excluded water from the active site pocket, we were interested to see if this was maintained when a new cavity in the pocket was introduced. The syringol-bound mutants, A, S, and V, contain an additional ordered water in the active site (Figure A2.6 of Appendix A2) which may help to maintain the substrate in a productive binding pose for catalysis. As expected, the bulkier GcoA-F169H mutant excludes water from the active site, as with the WT structure.

4.4.4 Syringol clashes with both GcoA-F169 and the substrate access lid in simulations of WT GcoA.

In the WT enzyme with syringol bound, active site crowding can be relieved in several ways. First, as already noted, syringol can shift towards the heme (Figure 4.2A); this effect is seen in MD simulations (80 ns) carried out on WT GcoA with either bound guaiacol or syringol, although the effect is much subtler than in the crystal structures (Figure A2.7 of Appendix A2). A second effect is more pronounced in MD simulations, which show that GcoA-F169 is significantly more flexible and perturbed from the crystal structure position when syringol is bound at the WT active site rather than guaiacol (Figure 4.3A, C, E, and Figure A2.8 of Appendix A2). This effect is complemented by the static picture given by the crystal structures (Figure 4.2A), which show GcoA-F169 is "pushed away" from the substrate by a distance commensurate with the observed shift of the substrate.

Opening of the substrate access loop, a larger scale phenomenon that is closely related to the movement of GcoA-F169, can also relieve active site crowding. All GcoA crystal structures to date present a closed active site "lid," but previous MD simulations demonstrated the ability of the F/G helices (and their connecting loop) to move away from the active site, thus exposing the active site to solution, particularly in the apo form.[70] Crystal packing may hinder lid opening in the GcoA-F169A structure; thus, the first two effects (shifting of substrate and GcoA-F169) are more pronounced in crystal structures. In MD of GcoA in solution, however, the active site loop is unconstrained, and the effect of the GcoA-F169 clash with syringol is observed less at the substrate and more on the enzyme. This includes the positioning and flexibility of GcoA-F169 (as mentioned above)

and the substrate access lid (Figure 4.3B, D, F, and Figure A2.9 of Appendix A2), which is significantly more prone to open with syringol bound in WT GcoA than with guaiacol, as well as either substrate bound in MD simulations on the GcoA-F169A mutant. Open and closed access loops have been observed in P450 enzymes P450cam[288] and BM3[289, 290] crystal structures. The open GcoA configurations we observe in MD simulations are about half as open than the aforementioned open crystal structures and possibly not sufficiently open to allow substrate ingress and egress. However, this principle observed over 80 ns is likely to be more pronounced over the course of the full catalytic cycle. We note as well that, to date, efforts to crystallize GcoA in the apo state have proven unsuccessful; when achieved, these may reveal a more open configuration of the substrate access lid.

The above conclusions from 80-ns MD simulations are also supported by a deeper analysis of three independent 1-μs MD trajectories of WT GcoA with syringol and guaiacol bound at the active site, which were originally presented in our previous study (Figure A2.21 of Appendix A2).[70] GcoA-F169 is significantly perturbed from its crystal structure position and more mobile; this coincides with an increased propensity to open the substrate access lid, which is the only region of significant difference in flexibility (Figure A2.10 of Appendix A2).

We also performed density functional theory (DFT) calculations on a truncated active site model, demonstrating that the *O*-demethylation of syringol proceeds via a similar pathway as previously described for guaiacol (Figure A2.11 and Table A2.6 of Appendix A2,).[70] Optimized transition state (TS) geometries and free energy barriers for the rate-limiting hydrogen atom transfer (HAT) are likewise very similar in the two cases. In addition, replica exchange thermodynamic integration (RETI) simulations were also

conducted to examine relative free energies associated with substrate binding and mutating GcoA-F169 in the closed state of the enzyme (Figure A2.12 and A2.13 of Appendix A2). The RETI simulations reveal additional substrate binding modes, made possible by the "softer" interactions between the substrate and enzyme, as TI simulations gradually "turn on" and "turn off" electrostatic and van der Waals components of intermolecular interactions, and quantify the effect of this binding flexibility on binding thermodynamics.

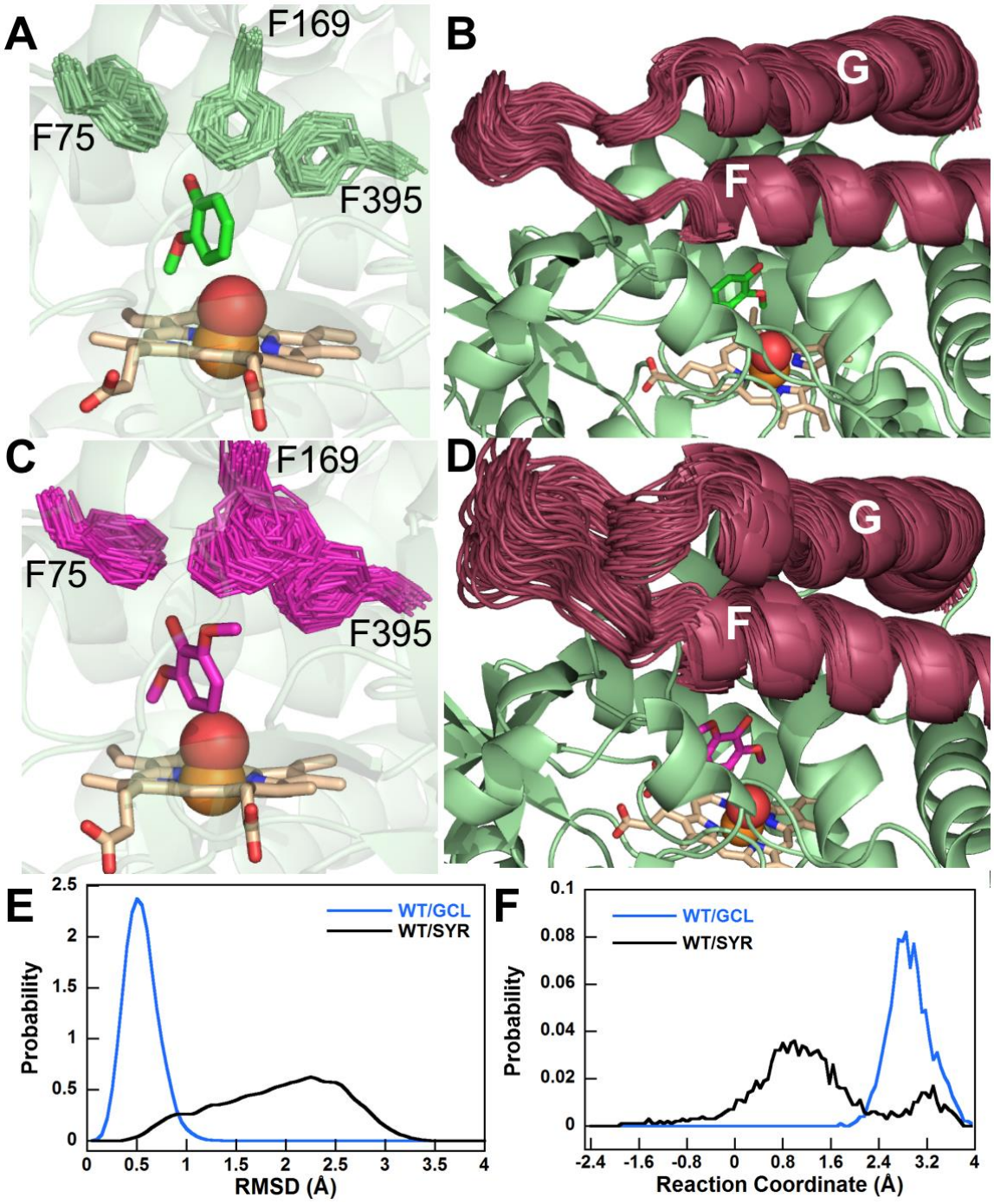


Figure 4.3 GcoA-F169 in WT GcoA and the substrate access loop are significantly displaced with bound syringol. MD simulations with bound guaiacol indicate that (A) GcoA-F169 and the (B) substrate access lid are relatively stable. Introducing syringol results in increased flexibility of (C) GcoA-F169 and (D) the substrate access loop. In A-

D, the position of each of the labeled Phe side chains (or alternatively the substrate access loop) is shown every 4 ns over the course of the 80 ns MD simulation. Substrate, the Phe side chains, and heme are shown in "sticks"; the Fe atom and the O atom of a reactive heme-oxo intermediate are shown as spheres. Probability distributions are shown for (E) the RMSD of the six ring carbons of GcoA-F169 from their crystal structure positions and (F) the reaction coordinate for opening/closing of the substrate access loop (as defined in Appendix A2; lower values indicate more open configurations).

4.4.5 Sequence position 169 in CYP255A enzymes is highly variable.

GcoA belongs to the CYP255A family of cytochrome P450 enzymes.[282] Conservation analyses of GcoA homologs revealed a notably variable 169 position among active site residues. Moreover, not only is GcoA-F169 the least conserved of the triad of phenylalanine residues in the active site, it is also among the least conserved positions in the entire protein (Figure 4.4, Figure A2.14 and Table A2.7 of Appendix A2). From a multiple sequence alignment, we determined that alanine and phenylalanine are the most frequent residues utilized by CYP255A enzymes at position 169 with alanine present in the majority of sequences. Hence, the GcoA-F169A mutant which showed enhanced turnover on guaiacol and syringol is closer to the CYP255A consensus protein than the WT. It is interesting that although none of the GcoA homologs in our analyses exhibits a histidine at position 169, the GcoA-F169H mutant was stable and showed the highest specific activity on guaiacol. Next to GcoA-F169, A295 and T296 respectively show the highest variability of residues in the active site. Besides these, other residues within 6 Å of the center of mass of the guaiacol substrate generally show high conservation.

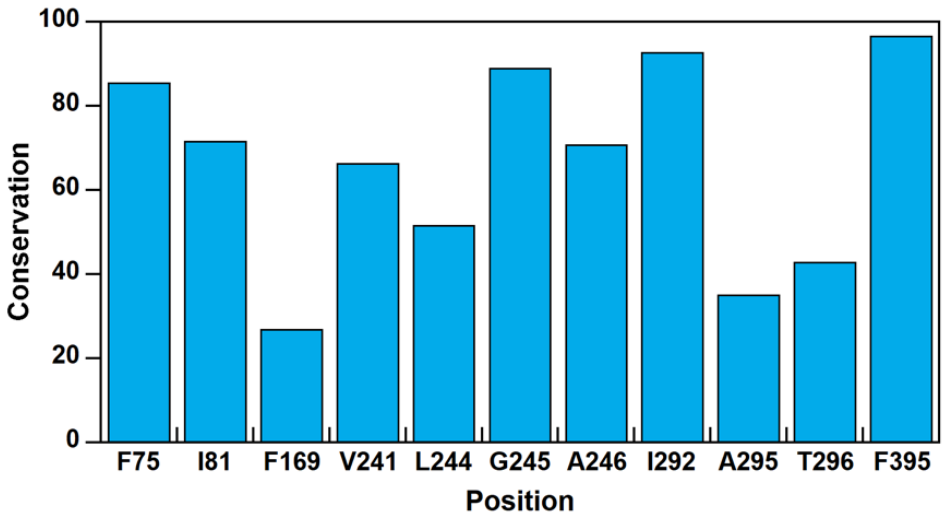


Figure 4.4 Bioinformatic analysis of CYP255A sequences indicates variability in the 169th sequence position.Conservation of residues within 6 Å of guaiacol in GcoA (PDB ID 5NCB), determined via an analysis of protein sequences from 482GcoA homologs. (See Appendix A2 for details.) Conservation scores are reported as percentiles. F169 is less conserved than 73% of positions in GcoA.

4.4.6 F169A enables in vivo syringol conversion by GcoA.

Finally, we aimed to demonstrate *in vivo* conversion of syringol to pyrogallol using *Pseudomonas putida* KT2440 because it possesses many native aromatic-catabolic pathways relevant to lignin conversion.[265, 284, 291, 292] To accomplish this, we transformed plasmids expressing WT GcoA or the GcoA-F169A variant and GcoB into a strain that constitutively overexpresses PcaHG, (Tables A2.8 and A2.9 of Appendix A2) a native 3,4-protocatechuate dioxygenase from *P. putida* that converts pyrogallol into 2-pyrone-6-carboxylate, a more stable product (Figure 4.5A).[293] When cultured with 20 mM glucose and 1 mM syringol, ^{1}H NMR analysis of the culture media revealed that peaks

corresponding to syringol completely disappeared in cells expressing GcoA-F169A (AM157) after 6 hours, which coincided with the appearance of 3MC, pyrogallol, and 2-pyrone 6-carboxylate (Figure 4.5B, Figure A2.15 of Appendix A2). A standard for 2-pyrone 6-carboxylate was not available; however, we did verify the presence of 2-pyrone 6-carboxylate using LC-MS-MS (Figure A2.15 of Appendix A2). While small amounts of pyrogallol, 3MC, and 2-pyrone 6-carboxylate were also observed with the WT GcoAB (AM156), nearly 60% of the original syringol remained after 6 hours. Pyrogallol or 3MC were not observed in the strain lacking GcoAB (AM155). These data indicate that the F169A variation enhances *in vivo O*-demethylation of syringol to 3MC and pyrogallol by GcoA.

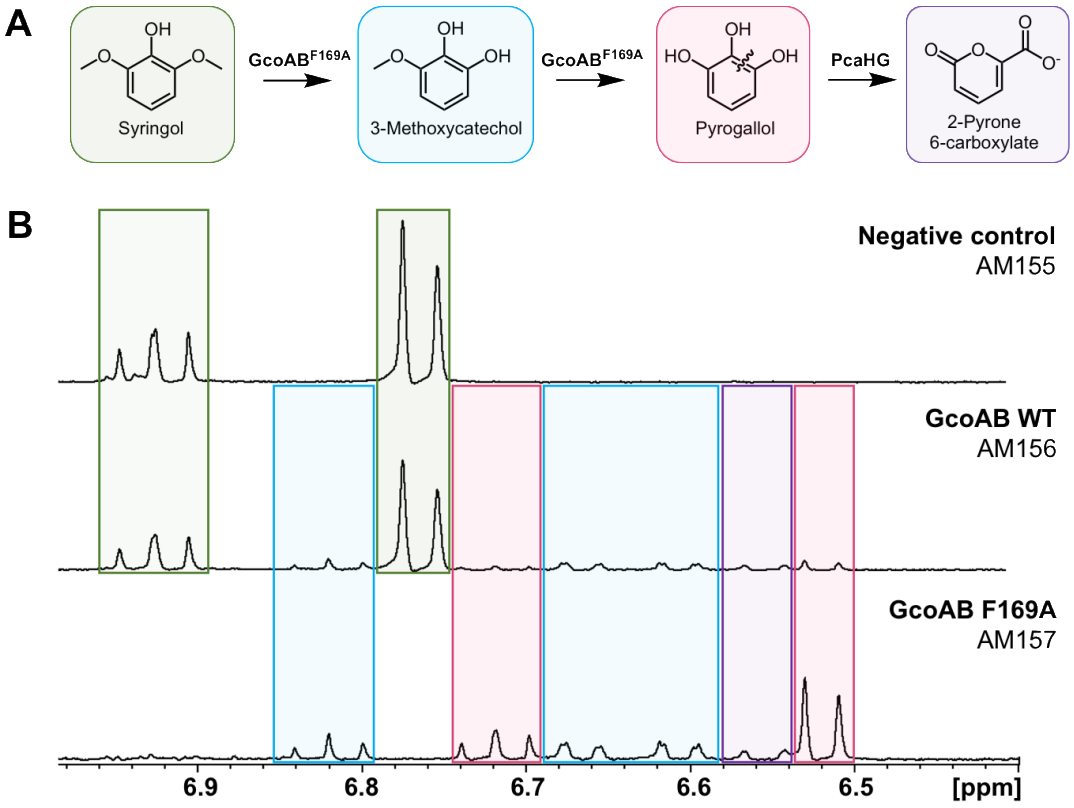


Figure 4.5 GcoA-F169A converts syringol in vivo.(A) A pathway for *in vivo* syringol *O*-demethylation to pyrogallol and cleavage to 2-pyrone 6-carboxylate is proposed. (B) After

6 hours, strains were analyzed for their ability to turn over syringol via ^{1}H NMR spectroscopy. Syringol (green) is completely converted to pyrogallol (pink), 3MC (blue), or 2-pyrone 6-carboxylate (purple) in AM157. The WT GcoA enzyme in AM156 showed only small amounts of conversion. AM155, which does not express GcoA, showed no conversion.

4.5 Discussion and conclusion

The creation of enzymes that overcome the challenge of lignin heterogeneity through increased substrate promiscuity is an attractive goal, but this might come at the cost of reduced activity towards the natural substrate. Unexpectedly, not only does GcoA-F169A bind both guaiacol and syringol in a productive orientation analogous to guaiacol in WT (Figure 4.2), but GcoA-F169A is also more catalytically efficient for *O*-demethylation of both guaiacol and syringol relative to WT, where *O*-demethylation is well-coupled to NADPH oxidation (Table 4.1, Figure 4.1). Alongside this biochemical observation, the bioinformatics analysis shows that alanine is the most prevalent residue in the 169th sequence position in the CYP255A family. Taken together, it is surprising that the WT GcoA does not possess an alanine at position 169 if guaiacol is the primary substrate, as assumed in the original reports of GcoAB.[70, 273, 282] Given that the GcoA-F169A mutation results in improved turnover of guaiacol, we speculate that either guaiacol is not the primary substrate or there has been little evolutionary pressure in *Amycolatopsis* sp. ATCC 39116 for improved turnover of guaiacol. Another potential explanation could be that syringol *O*-demethylation and subsequent ring cleavage leads to dead-end products

that cannot be catabolized and may be toxic to the microbe; thus, GcoA-F169 could function to prevent natural syringol catabolism.

As a first step towards enabling syringol catabolism, *in vivo* experiments validated the *in vitro* studies by illustrating efficient *O*-demethylation of syringol and 3MC by the GcoA-F169 mutant. While 2-pyrone 6-carboxylate was detected, pyrogallol is a poor substrate of PcaHG, as most of the intermediate is lost to oxidation (Figure A2.15 of Appendix A2), and it is unclear whether 2-pyrone-6-carboxylate can be catabolized further. Future work could focus on identifying or developing a dioxygenase capable of cleaving pyrogallol to a product that could be further metabolized. Coupled with GcoAB, the *meta*-cleavage pathway of *P. putida* mt-2 might enable complete assimilation of syringol if pyrogallol can be efficiently cleaved to 2-hydroxymuconate.

Cytochrome P450 systems are one of the most versatile classes of enzymes, making them an ideal target for engineering enhanced activity and substrate promiscuity. Our system is a prime example. The mutation of a single residue resulted in efficient turnover of an S-lignin substrate, syringol, in addition to the native G-subunit substrate, guaiacol, which is not efficiently achieved in the WT enzyme. The plasticity of the GcoA active site may be amenable to yet more modifications, allowing us to encompass other lignin monomers as substrates, such as 4-substituted compounds (e.g., vanillin, syringaldehyde). Indeed, in previous work we showed that T296 sterically clashes with the C4 position of vanillin, preventing *O*-demethylation.[70] Interestingly, the 296 position is also quite variable according to the bioinformatics analysis (Figure 4.4). The results described here suggest that we may employ a similar structure-guided approach to investigate the activity of several T296 variants on 4-substituted compounds. As a large number of lignin degradation

products contain *para*-substituted R-groups that are bulkier than hydrogen, creating an engineered bacterium where a minimal number of genetic insertions leads to maximal lignin bioconversion is of keen interest for future work. More broadly, the evolutionary trajectory, substrate specificity, and catalytic efficiency of the CYP255A family of aromatic *O*-demethylases will be the subject of future work to elucidate the relevance of this cytochrome P450 family for microbial lignin conversion.

4.6 Methods

4.6.1 Protein expression and purification.

Mutagenesis was performed using primers listed in Appendix A2, with the Q5 polymerase and KLD enzyme mix (NEB) according to the manufacturer's protocol. Proteins were expressed as previously described.[70]

4.6.2 Crystallization and structure determination.

Crystallization, diffraction experiments, and structure solution were carried out as previously described.[70]

4.6.3 Biochemical characterization

Heme quantification of GcoA-F169 mutants. Catalytically active heme bound to each GcoA mutant was determined as previously described.[70, 294] Detailed methods are provided in Appendix A2.

Determination of [FAD] and non-heme [Fe] in GcoB. The FAD and 2Fe-2S content of GcoB was measured as previously described.[70, 295] Detailed methods are given in Appendix A2.

Steady state kinetics of GcoA-F169A. The *O*-demethylation reactions of guaiacol, syringol, and 3MC were continuously monitored using the NADH consumption assay as previously described.[70] Detailed methods are provided in Appendix A2.

Determination of substrate dissociation constants (K_D) with GcoA-F169A. The equilibrium binding constant, K_D, for GcoA-F169A and guaiacol, syringol, and 3MC was determined as previously described.[70] See Appendix A2 for detailed methods.

Fomaldehyde Determination. The [formaldehyde] produced upon reaction with GcoA-F169A and substrates was determined using a colorimetric assay with tryptophan.[70, 296] See Appendix A2 for detailed information.

HPLC for product identification and specific activity measurement. HPLC was used to verify the *O*-demethylated product of GcoA-F169A GcoA/GcoB with guaiacol, syringol, or 3MC. In addition, discontinuous HPLC was used to determine the specific activity of aromatic product disappearance. For detailed methods, see Appendix A2.

Detection of H_2O_2 via HRP and Amplex Red assay. A colorimetric assay involving horseradish peroxidase (HRP) and Amplex Red was used to quantify H_2O_2 in the reaction between GcoA-F169A GcoA/GcoB, NADH and guaiacol, syringol, or 3MC. Detailed methods are given in Appendix A2.

4.6.4 Molecular dynamics, density functional theory, and bioinformatics.

MD simulations and DFT calculations were performed with similar methodology as in our previous work.[70] Full details of the computational methods and references can be

found in Appendix A2. 482 homologous CYP255A sequences were retrieved from a BLASTP search against GcoA. After multiple sequence alignment (MSA), conservation was analyzed from relative entropy calculations for each site. Further details can be found in Appendix A2.

4.6.5 In vivo syringol utilization.

Strains used for shake flask experiments were grown overnight in LB media and resuspended the following day in M9 minimal media with 20 mM glucose as described in the Appendix A2. Cells were grown until they reached an OD_{600} of ~1, at which point syringol was added at a final concentration of 1 mM. ^{1}H NMR spectroscopy was used to analyze syringol consumption.

CHAPTER 5. Characterization of a Two-Enzyme System for Plastics Depolymerization

This chapter has been reprinted with permission from Knott et al.,[297] Copyright 2020, National Academy of Sciences USA. The author of this dissertation performed the phylogenetic analysis and bioinformatic analysis of MHETase active site and key residues, which provided a groundwork for further structural, biochemical, and computational studies of MHETase. Collaborators at the National Renewable Energy Laboratory (Brandon C. Knott, Erika Ericson, Isabel Pardo, Jared J. Anderson, Graham Dominick, Christopher W. Johnson, Nicholas A. Rorrer, Caralyn J Szostkiewicz, and Bryon S. Donohoe), University of Portsmouth (Mark D. Allen, Rosie Graham, and Harry P. Austin), University of Florida (Fiona L. Kearns and H. Lee Woodcock), and Montana State University (Ece Topuzlu and Valérie Copié) performed other parts of the study.

5.1 Abstract

Plastics pollution represents a global environmental crisis. In response, microbes are evolving the capacity to utilize synthetic polymers as carbon and energy sources. Recently, *Ideonella sakaiensis* was reported to secrete a two-enzyme system to deconstruct polyethylene terephthalate (PET) to its constituent monomers. Specifically, the *I. sakaiensis* PETase depolymerizes PET, liberating soluble products including mono-(2-hydroxyethyl) terephthalate (MHET), which is cleaved to terephthalic acid and ethylene glycol by MHETase. Here, we report a 1.6 Å resolution MHETase structure, illustrating that the MHETase core domain is similar to PETase, capped by a lid domain. Simulations

of the catalytic itinerary predict that MHETase follows the canonical two-step serine hydrolase mechanism. Bioinformatics analysis suggests that MHETase evolved from ferulic acid esterases, and two homologous enzymes are shown to exhibit MHET turnover. Analysis of the two homologous enzymes and the MHETase S131G mutant demonstrates the importance of this residue for accommodation of MHET in the active site. We also demonstrate that the MHETase lid is crucial for hydrolysis of MHET and, furthermore, that MHETase does not turnover mono-(2-hydroxyethyl)-furanoate or mono-(2-hydroxyethyl)-isophthalate. A highly synergistic relationship between PETase and MHETase was observed for the conversion of amorphous PET film to monomers across all non-zero MHETase concentrations tested. Lastly, we compare the performance of MHETase:PETase chimeric proteins of varying linker lengths, which all exhibit improved PET and MHET turnover relative to the free enzymes. Together, these results offer insights into the two-enzyme PET depolymerization system and will inform future efforts in the biological deconstruction and upcycling of mixed plastics.

5.2 Significance

Deconstruction of recalcitrant polymers such as cellulose or chitin is accomplished in nature by synergistic enzyme cocktails that evolved over millions of years. In these systems, soluble dimeric or oligomeric intermediates are typically released via interfacial biocatalysis, and additional enzymes often process the soluble intermediates into monomers for microbial uptake. The recent discovery of a two-enzyme system for PET deconstruction, which employs one enzyme to convert the polymer into soluble intermediates (PETase) and another enzyme to produce the constituent PET monomers

(MHETase), suggests that nature may be evolving similar deconstruction strategies for synthetic plastics. The current study on the characterization of the MHETase enzyme and synergy of the two-enzyme PET depolymerization system may inform enzyme cocktail-based strategies for plastics upcycling.

5.3 Introduction

Synthetic polymers pervade all aspects of modern life, due to their low cost, high durability, and impressive range of tunability. Originally developed to avoid the use of animal-based products, plastics have now become so widespread that their leakage into the biosphere and accumulation in landfills is creating a global-scale environmental crisis. Indeed, plastics have been found widespread in the world's oceans,[298-301] in the soil,[302] and more recently, microplastics have been observed entrained in the air.[303] The leakage of plastics into the environment on a planetary scale has led to the subsequent discovery of multiple biological systems able to convert man-made polymers for use as a carbon and energy source.[72, 304-308] These plastic-degrading systems offer a starting point for biotechnology applications towards a circular materials economy.[78, 308-311]

Among synthetic polymers manufactured today, polyethylene terephthalate (PET) is the most abundant polyester, which is made from petroleum-derived terephthalic acid (TPA) and ethylene glycol (EG). Given the prevalence of esterase enzymes in nature, PET biodegradation has been studied for nearly two decades, with multiple cutinase enzymes reported to perform depolymerization.[79, 312-320] In 2016, Yoshida *et al*. reported the discovery and characterization of the soil bacterium, *Ideonella sakaiensis* 201-F6, which employs a two-enzyme system to depolymerize PET to TPA and EG, which are further

catabolized as a carbon and energy source.[72] Characterization of *I. sakaiensis* revealed the PETase enzyme, which is a cutinase-like serine hydrolase that attacks the PET polymer, liberating bis-(hydroxyethyl) terephthalate (BHET), mono-(2-hydroxyethyl) terephthalate (MHET), and TPA. PETase cleaves BHET to MHET and EG, and the soluble MHET product is further hydrolyzed by MHETase to produce TPA and EG. Multiple crystal structures and biochemical studies of *I. sakaiensis* PETase revealed an open active site architecture that is able to bind to PET oligomers.[81, 82, 321-325] The PETase enzyme likely follows the canonical serine hydrolase catalytic mechanism,[326] but open questions remain regarding the mobility of certain residues during the catalytic cycle.[321]

Conversely, the structure and function of the MHETase enzyme is far less characterized, with only two published studies focused on MHETase structure and engineering to date.[85, 86] These studies report structures at 2.1-2.2 Å resolution, wherein the similarity to ferulic acid esterase (FAE) is noted.[327, 328] Informed by these structures, engineering efforts aimed to improve turnover of BHET by MHETase, which is a non-native substrate of the wild-type. Beyond these studies and the original report of MHETase from Yoshida *et al.*,[72] several questions remain regarding the MHETase mechanism and PETase-MHETase synergy. To that end, here we combine structural, computational, biochemical, and bioinformatics approaches to reveal molecular insights into the MHETase structure, mechanism of hydrolysis, the evolution of MHETase activity from FAEs, and engineering of the two-enzyme system for PET depolymerization.

5.4 Results

5.4.1 Structural characterization of MHETase reveals a core domain similar to that of PETase.

Four crystal structures of MHETase were obtained with the highest resolution data (6QZ3) extending to 1.6 Å with a benzoate molecule in the active site (Figure 5.1, Table A3.1 of Appendix A3). These data reveal a catalytic domain that adopts the α/β-hydrolase fold typical of a serine hydrolase, with an extensive lid domain (Figure 5.1A), that partially covers the active site and hosts a well-coordinated Ca^{2+} cation. A similar Ca^{2+} binding site was characterized for *Aspergillus oryzae* FaeB (*Ao*FaeB), wherein it was hypothesized to have a role in stabilizing the lid domain.[328] Overall, the structure of MHETase is most similar to that of FAEs, as discussed previously.[85, 86] The structural conservation between the hydrolase domains of MHETase and PETase is striking (Figure 5.1D and Figure A3.1 of Appendix A3), and despite the large insertion of ~240 residues representing the lid domain, residues Ser225, Asp492, and His528 effectively reconstitute the catalytic triad (Figure 5.1B). In fact, the terminal residues of the lid domain converge to within hydrogen-bonding distance of each other (Tyr252-Ala469, 2.9 Å), creating a compact linkage to the hydrolase domain. The lid domain of MHETase is exceptionally large, as average lid domains in α/β-hydrolases tend to be ~100 residues,[326] and is more typical of a lid from tannase family members (*vide infra*). The equivalent connection site in PETase is occupied by a seven-residue loop.

In addition, we determined two apo structures with alternative packing (6QZ2 and 6QZ4), one structure with a fully occupied benzoic acid ligand (6QZ3), and one with partially occupied benzoic acid (6QZ1). We observed that residue Phe415 adopts a ‘closed’

orientation on substrate binding consistent with prior substrate bound structures (PDB IDs 6QGA and 6QGB),[85] and the partially occupied site results in an intermediate dual 'open/closed' conformation (Figure A3.2A-C of Appendix A3).[85] The only other amino acid with side chain positioning correlated with ligand binding is Gln410. When Phe415 is in the open position, the side chain of Gln410 pivots toward the active site to a position wherein the heavy atoms would be as close as 1.8 Å to those of Phe415 if it were in the closed conformation (Figure A3.2D of Appendix A3).

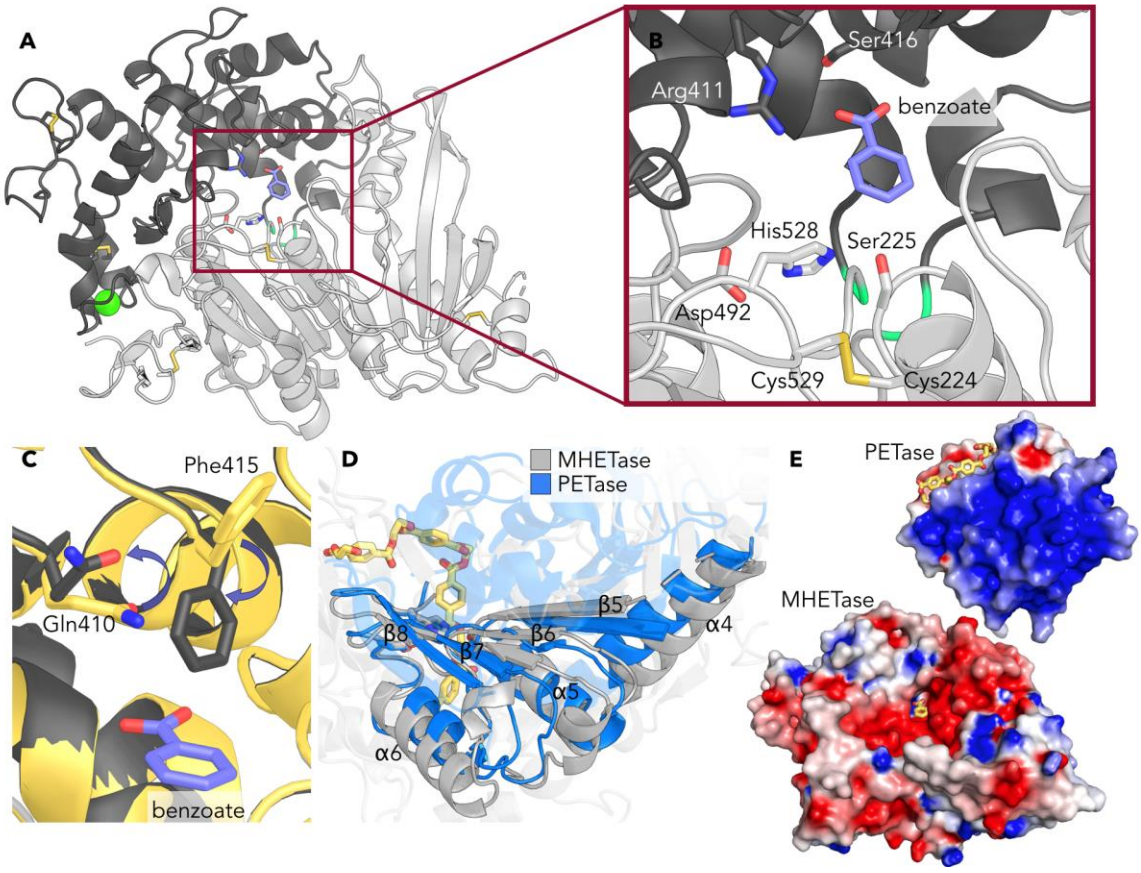


Figure 5.1 MHETase structural analysis.(A) MHETase structure (1.6 Å resolution, PDB code 6QZ3) highlighting the catalytic triad, five disulfides (in yellow and gray stick representation), benzoate (purple sticks), and calcium ion (green sphere). The lid domain is dark gray, whereas the hydrolase domain is light gray. Main chain atoms of the linkage

residues Tyr252 and Ala469 are colored lime green (also in panel B). (B) Closeup of the MHETase active site with benzoate bound; catalytic triad, active site disulfide, Ser416, and Arg411 shown as sticks. (C) The concerted movement of residues Gln410 and Phe415 on ligand binding is illustrated with purple arrows in a superposition of the apo enzyme (yellow) with the ligand bound state (gray). The relative position of benzoic acid is depicted in purple. (D) Structural comparison between MHETase (gray) and PETase (PDB code 6EQE, in blue), highlighting regions of alignment in the hydrolase domain. A PET tetramer from a prior docking study is shown in yellow sticks (also in panel E).[81] (E) Electrostatic potential distribution mapped to the solvent-accessible surface of PETase [81] and MHETase as a colored gradient from red (acidic) at -7 *kT/e* to blue (basic) at 7 *kT/e* (where *k* is the Boltzmann's constant, *T* is temperature, and *e* is the charge of an electron). PETase is shown with a bound PET tetramer, and MHETase with benzoate bound from the 6QZ3 structure (yellow). The models are drawn to scale and aligned via their catalytic triad demonstrating their relative size difference.

Given the difference in overall isoelectric point (pI) between PETase (9.65) and MHETase (5.11), we generated electrostatic surface profiles for comparison (Figure 5.1E). As previously reported, PETase has a highly polarized surface charge,[81] whereas MHETase exhibits a more heterogeneous and acidic surface charge distribution. MHETase contains five disulfide bonds (Figure 5.1A). One of the MHETase disulfides is located at the active site, connecting cysteines (Cys224 and Cys529) adjacent to the catalytic residues (Ser225 and His528, respectively), which is conserved in tannase family members [328]. PETase lacks a structurally equivalent disulfide, and the aligning residues in PETase (Trp159 and

Ser238) are the same two residues mutated by Austin *et al.* to yield a PETase substrate binding groove similar to that of cutinases, resulting in improved activity on a crystalline PET substrate.[81]

5.4.2 Molecular simulations of the MHETase reaction predict deacylation is rate-limiting.

The MHETase structure suggests a serine hydrolase mechanism for MHET hydrolysis.[326] To elucidate the detailed reaction mechanism, we first constructed a Michaelis complex *in silico* utilizing the CHARMM molecular simulation package (details in Appendix A3).[329] To examine MHETase dynamics and ligand stability, classical molecular dynamics (MD) simulations were conducted with NAMD[330] (all simulations totaling 2.25 μs) utilizing the CHARMM forcefield.[331] Given the observed dual occupancy for Phe415 positioning in the crystal structures, we simulated in triplicate (each simulation 150 ns in length) the four combinations of Phe415 position ("open" and "closed") and active site occupancy (empty active site and MHET-bound). In each case wherein Phe415 begins in the closed state (starting configurations from 6QZ3 structure, with coordinates for residues 56-61 from 6QZ4), Phe415 opens in the first 10 ns and rarely returns to the closed state; simulations that begin with Phe415 open (built from 6QZ4 structure) all remain open. To examine the effect of calcium binding, a fifth scenario absent of either MHET or Ca^{2+} was modeled in triplicate 150 ns simulations (the prior four scenarios each include bound Ca^{2+}). These trajectories show evidence for lid stabilization upon Ca^{2+} binding mainly in the immediate vicinity of the calcium binding site (Figure A3.3 of Appendix A3). When bound at the active site, the carboxylate motif of MHET exhibits

stable hydrogen bonds with Arg411 and Ser416, while the carbonyl is stabilized via hydrogen bonds to the oxyanion hole residues, Glu226 and Gly132 (Figure A3.4 and A3.5 of Appendix A3). In all three simulations with MHET bound and Phe415 open, MHET maintains these interactions and remains bound at the active site throughout the duration of the 150 ns simulation and primed for hydrolysis (hydrogen bond distance between Ser225 and His528=2.0±0.2 Å; nucleophilic attack distance between Ser225 and MHET=3.1±0.3 Å; hydrogen bond distance between Asp492 and His528=1.8±0.1 Å). Further analysis of the MD simulations, including time traces for these important interactions, is available in the Appendix A3.

Serine hydrolases catalyze a two-step reaction involving formation of an acyl-enzyme intermediate (acylation) that is released hydrolytically in the second step (deacylation).[326] We utilized the Amber software package[332] to perform hybrid quantum mechanics/molecular mechanics (QM/MM) two-dimensional umbrella sampling with semi-empirical force field SCC-DFTB[333] to study the catalytic steps. Judicious selection of a reaction coordinate is critical for kinetically meaningful barrier calculations. We chose the forming and breaking C-O bonds to map the free energy landscape for both reaction steps, informed by transition path sampling studies of other hydrolase enzymes.[334, 335]

In acylation, the catalytic serine (Ser225) is deprotonated by His528, activating it for nucleophilic attack upon the carbonyl C of MHET, liberating EG and forming the acyl-enzyme intermediate (AEI, Figure 5.2A-C). The minimum free energy path (MFEP) computed from QM/MM 2D-umbrella sampling (along the forming C-O bond between the MHET carbonyl C and Ser225 and the breaking MHET C-O ester bond) predicts an acylation free energy barrier ($\Delta G^{\ddagger}$) of 13.9 ± 0.17 kcal/mol with an overall reaction free

energy ($\Delta G_{reaction}$) of -5.2 ± 0.04 kcal/mol (Figure 5.2G, Figure A3.6A of Appendix A3). Although serine hydrolases have at times been considered to proceed through metastable tetrahedral intermediates along the reaction pathway for acylation and deacylation,[336-338] the acylation MFEP calculated from 2D umbrella sampling does not indicate intermediate configurations with metastability.

Following cleavage of EG from MHET, classical MD simulation reveals that EG leaves the active site in the presence of the AEI (Figure 5.2H). In one simulation, EG initially maintains a hydrogen bond with His528 for ~100 ps, then dislodges from the active site, and is free in solution within 1 ns. Three identical simulations were initiated, and EG exits the active site within 4 ns in each. An important implication of this observation is that the deacylation reaction proceeds without EG in the active site. This allows greater access for water molecules to approach the charged nitrogen of His528 for deacylation (Figure A3.7 of Appendix A3).

Deacylation involves nucleophilic attack by a water molecule on the AEI, liberating TPA (Figure 5.2D-F). His528 plays a similar role as in acylation, deprotonating the catalytic water and transferring this proton to the catalytic serine, regenerating Ser225 for another catalytic cycle. The MFEP computed from QM/MM 2D-umbrella sampling (along the forming C-O bond between MHET and water and the breaking AEI C-O bond) reveals a deacylation free energy barrier ($\Delta G^{\ddagger}$) of 19.8 ± 0.10 kcal/mol and an overall reaction free energy ($\Delta G_{reaction}$) of 2.6 ± 0.07 kcal/mol (Figure 5.2I, Figure A3.6B of Appendix A3). Together, the two catalytic steps are exergonic by -2.6 ± 0.08 kcal/mol. Deacylation is predicted to be the rate-limiting step, with a rate of $7.1 \pm 1.1 \times 10^{-2}$ s^{-1} (from transition state theory, at 30°C, and assuming a transmission coefficient of 1), more than four orders of

magnitude slower than acylation ($1.02 \pm 0.28 \times 10^3$ s^{-1}). As in acylation, metastable configurations along the MFEP are not observed.

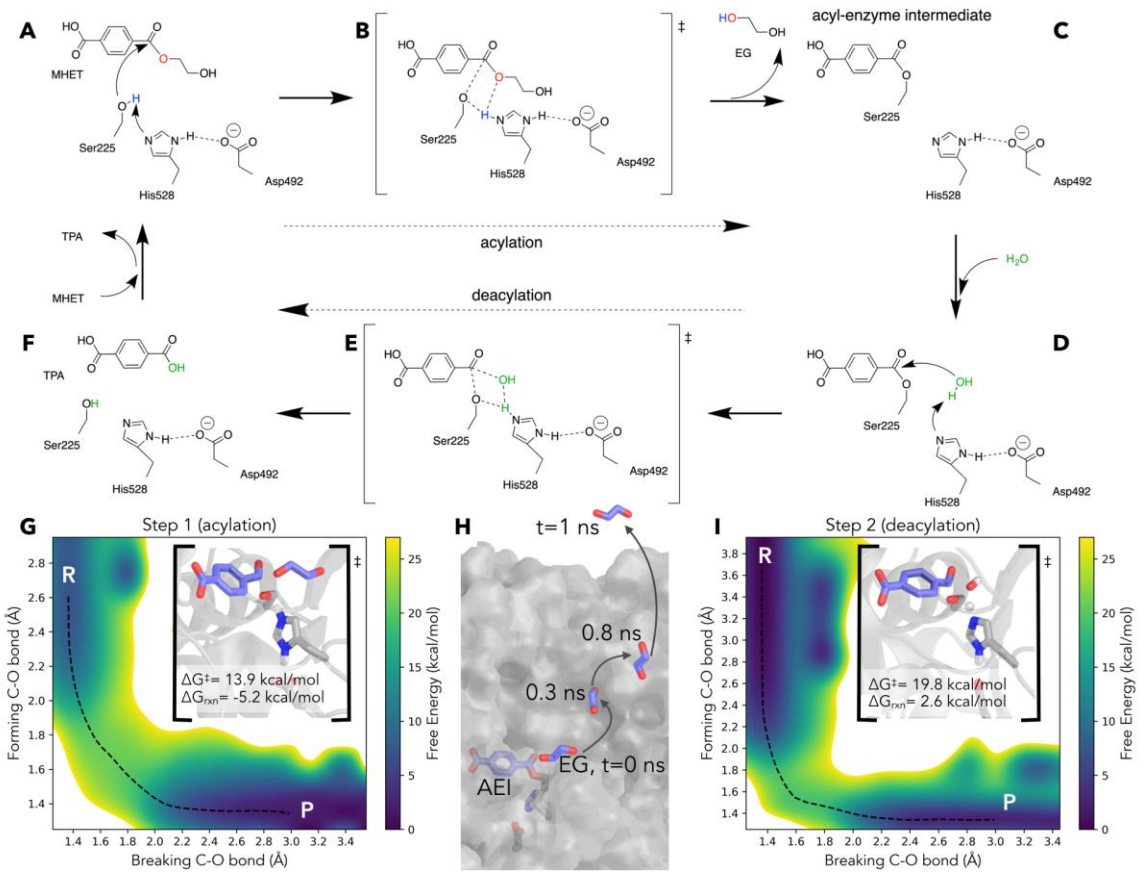


Figure 5.2 The MHETase catalytic mechanism. (A) reactant, (B) transition state, and (C) product of acylation in which His528 transfers a proton from Ser225 to the ethylene glycol (EG) leaving group. In deacylation (panels D-F), His528 plays a similar role and restores the catalytic serine, transferring a proton from a water molecule to Ser225 and generating a free terephthalic acid (TPA) molecule. (G) The free energy surface for acylation computed along a reaction coordinate described by the breaking and forming C-O bonds. The minimum free energy path is shown in black dashes. H) Following acylation, EG leaves the active site within 1 ns of a classical MD simulation. (I) The free energy surface

for deacylation, exhibiting a predicted higher barrier than acylation. The minimum free energy path is shown in black dashes.

5.4.3 Bioinformatics analysis suggests that MHETase evolved from a ferulic acid esterase.

Beyond structural and mechanistic investigations, we were also interested in understanding potential MHETase evolutionary ancestry and identifying other MHET-active enzymes from natural diversity. MHETase belongs to the tannase family (PFAM ID: PF07519), which consists of fungal and bacterial FAEs, fungal and bacterial tannases, and several bacterial homologs of unknown function.[339] To elucidate sequence relationships between MHETase and tannase family enzymes, we performed bioinformatic analyses of 6,671 tannase family sequences retrieved from NCBI via PSI-BLAST.[340] MHETase shares low sequence similarity (<53%) with most sequences in the family, with the exception of homologs from *Comamonas thiooxydans* strains DS1, DF1 and DF2 (strain: NCBI:txid363952, protein:Genbank WP_080747404.1)[341] and *Hydrogenophaga* sp. PML113 (strain: NCBI:txid1899350, protein:Genbank WP_083293388.1), which exhibit 81% and 73% identity to MHETase, respectively (Figure 5.3A). Since initial identification of the homologous *C. thiooxydans* sequence (WP_080747404.1), this entry was removed from Genbank, as discussed in Appendix A3.

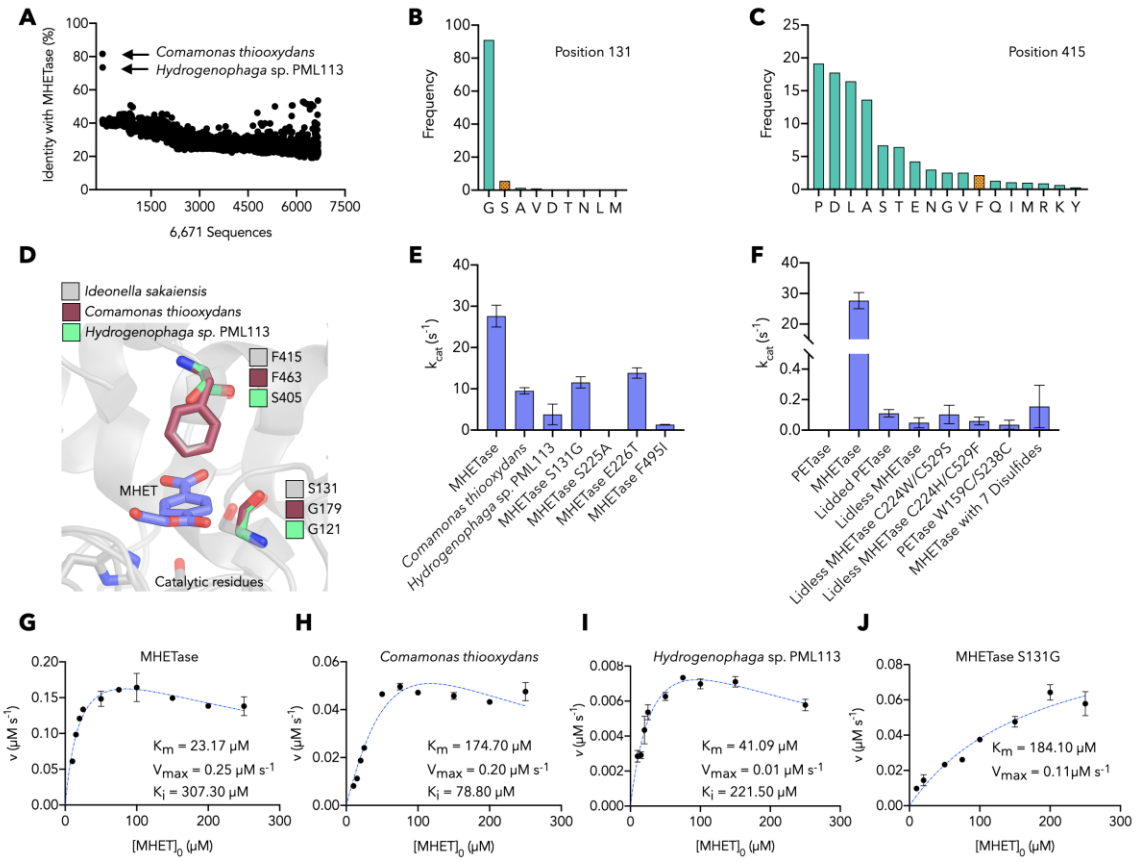


Figure 5.3 Characterization of MHETase, homologs, and mutants. (A) Sequence identity of 6,671 tannase family sequences retrieved by PSI-BLAST compared to MHETase. Sequences (*x*-axis) are in the same order returned by PSI-BLAST. Panels B-C show conservation analysis of residue positions 131 (panel B) and 415 (panel C) (using MHETase numbering). Frequency of each amino acid is based on a multiple sequence alignment of the 6,671 tannase family sequences. The residue found in MHETase at each position is indicated in orange. (D) Homology model of the MHET-bound active site within 6 Å of the bound substrate comparing MHETase to homology models of the *C. thiooxydans* and *Hydrogenophaga* sp. PML113 homologs (generated by SWISS-MODEL),[342] showing sequence variation at residue positions corresponding to Ser131 and Phe415 in MHETase. Panels E-F show the rate of enzymatic turnover of MHET determined for MHETase, both

homologous enzymes, and the indicated MHETase mutants, all of which are active on MHET save the catalytic mutant (S225A)(panel E), and enzymatic turnover rates for PETase, MHETase, and selected mutants on MHET (panel F), using 5 nM purified enzyme and 250 μM substrate at 30°C. Panels G-J show the initial enzyme reaction velocity as a function of substrate concentration for MHETase, *C. thiooxydans*, *Hydrogenophaga* sp. PML113, and the MHETase S131G mutant, respectively. Dashed blue lines represent the Michaelis-Menten kinetic model fit with substrate inhibition (panels F-H) or fit with the simple Michaelis-Menten model (panel I). Key kinetic parameters are provided in the inset. Additional parameters and confidence intervals on the listed parameters are provided in Table A3.3 of Appendix A3.

Using the multiple sequence alignment of 6,671 tannase family sequences, we performed conservation analysis with MHETase sequence positions as a reference (Figure A3.8 and A3.9 of Appendix A3), which shows that most positions in the active site are highly conserved. Notable exceptions are positions 257, 411, 415, and 416, which exhibit low conservation scores and are less conserved than 80% of MHETase positions overall (Figure A3.8B-C of Appendix A3). It is noteworthy (*vide infra*) that position 131 is a well-conserved glycine in 91% of tannase family sequences but serine appears at position 131 in MHETase (Figure 6.3B). Furthermore, the ten cysteine positions in MHETase that form five disulfide bonds are highly conserved in the tannase family (Figure A3.10A of Appendix A3). Although a sixth disulfide bond exists in *Ao*FaeB,[328] less than 8% of tannase family sequences exhibit this sixth disulfide bond, and the sixth disulfide bond positions are variable among this set (Figure A3.10B of Appendix A3). One cysteine of the sixth

disulfide bond in *Ao*FaeB is a single residue variation found in MHETase,[328] whereas the other sits on a loop where a 15-residue deletion is found in MHETase.

With this large dataset, we further conducted phylogenetic analysis of 120 sequences selected from tannase family sequences that were clearly annotated as tannases or FAEs in GenBank, including MHETase (Table A3.2 of Appendix A3). In the phylogenetic tree (Figure A3.11 of Appendix A3), bacterial and fungal enzymes form paraphyletic groups, and within these groups, there are separate FAE and tannase sub-groups. MHETase and the *C. thiooxydans* and *Hydrogenophaga* sp. PML113 homologs are found within a group of proteobacterial FAEs (bootstrap value>95%). In addition, when separate profile hidden Markov models (pHMM) are constructed with the annotated tannase family FAE and tannase sequences,[229] and then aligned with MHETase, a higher alignment score is achieved with the FAE pHMM than with the tannase pHMM (456.8 vs. 396.8), suggesting that MHETase is more similar to FAEs than tannases.

5.4.4 Biochemical characterization of active-site MHETase mutants and homologs reveals important residues for MHET hydrolytic activity.

From the bioinformatics analyses, we selected the MHETase homologs from *C. thiooxydans* and *Hydrogenophaga* sp. PML113 (Figure 5.3A) to test for MHET hydrolysis activity, which, along with the *I. sakaiensis* MHETase, were produced in *Escherichia coli* and purified. Activity assays were performed for each enzyme to determine MHET turnover rates (Figure 5.3E). The turnover rate (k_{cat}) for MHETase is 27.6±2.6 s^{-1}, as compared to 9.5±0.8 s^{-1} and 3.8±2.5 s^{-1} for the *C. thiooxydans* and *Hydrogenophaga* sp. PML113 enzymes, respectively. The enzymes were also evaluated over a range of substrate

concentrations to determine the Michaelis-Menten kinetic parameters (Figure 5.3G-J, Table A3.3 of Appendix A3). FAEs have been shown to exhibit concentration-dependent substrate inhibition,[343] in addition to the likely product inhibition of the enzyme.[85] MHETase and both homologs also display this behavior. Using a substrate inhibition model (details in Appendix A3), evaluation of the substrate-dependent reaction kinetics shows that MHETase more efficiently accepts MHET as a substrate than the *C. thiooxydans* and *Hydrogenophaga* sp. PML113 homologs, demonstrated by a K_m value of 23.17±1.65 μM as compared to values of 174.70±4.75 μM and 41.09±3.38 μM, respectively (Figure 5.3G-I, Table A3.3 of Appendix A3). However, MHETase is also the most susceptible to substrate inhibition with a K_k value of 307.30±20.65 μM. Despite the difference in affinity for MHET, MHETase and the *C. thiooxydans* enzyme exhibit similar maximal reaction rates, while the enzyme from *Hydrogenophaga* sp. PML113 is slower. The MHETase reaction efficiency, reported as k_{cat}/K_m,is ~10-fold higher than for the *C. thiooxydans* enzyme and ~20-fold higher than the *Hydrogenophaga* sp. PML113 enzyme.

Homology models of both the *C. thiooxydans* and *Hydrogenophaga* sp. PML113 enzymes were constructed with SWISS-MODEL[342] using the MHETase structure as a template (PDB ID 6QZ3), and the active site aligned with a modeled MHET-bound MHETase structure (Figure 5.3D). As noted, position 131 is a serine in MHETase, but a glycine in the two homologs (*C. thiooxydans*, Gly179 and *Hydrogenophaga* sp. PML113, Gly121) (Figure 5.3B). The *C. thiooxydans* enzyme is otherwise identical within 6 Å of the docked MHET ligand, whereas the *Hydrogenophaga* sp. PML113 enzyme also exhibits a serine in the equivalent position to the MHETase residue Phe415 (Figure 5.3C). An S131G mutant of MHETase was constructed to examine the role of this residue in MHET

hydrolytic activity, and steady-state enzyme kinetics were evaluated. The MHETase S131G mutant does not demonstrate concentration-dependent substrate inhibition as is observed for the wild-type enzyme, which is likely due to the poor affinity for MHET. The S131G mutant has a K_m value ~8-fold higher than that of wild-type MHETase and the reaction efficiency is reduced to ~3% that of the wild-type (Figure 5.3J, Table A3.3 of Appendix A3), illustrating the importance of this residue in MHET turnover.

Focusing on residues within the coordination sphere of the docked MHET ligand, amino acids and their frequencies across the tannase family were compared to MHETase (Figure. A3.8 and A3.9 of Appendix A3). In position 495, MHETase features a phenylalanine, while isoleucine is also a common residue in this position across the tannase family. Palm *et al.* demonstrated that Phe495 has a modest effect on activity by mutation to alanine.[85] We constructed and evaluated an MHETase mutant with isoleucine in this position (F495I), which dramatically impairs activity, lowering the turnover rate from 27.6±2.6 s^{-1} to 1.3±0.7 s^{-1} (Figure 5.3E). In position 226, which is part of the conserved "lipase box" motif in serine hydrolases,[344] MHETase exhibits a glutamate, while threonine and asparagine are more common amongst tannase family members. Mutation of this lipase box residue to threonine (E226T) yielded a ~50% reduction in MHET activity relative to the wild-type MHETase (Figure 5.3E). Mutation of the catalytic serine (S225A), as expected, produced an inactive enzyme.

5.4.5 Unique structural features between MHETase and PETase determine substrate specificity and stability.

Given the structural similarities of the MHETase and PETase core domains (Figure 5.1C-D), we were interested in understanding the role of unique MHETase features,

namely the lid domain and the active site disulfide bond between Cys224 and Cys529, on substrate specificity and MHET hydrolytic activity. Accordingly, the lid was both added to PETase ("lidded PETase") and removed from MHETase ("lidless MHETase"). Given the natural substrate specificities of wild-type PETase and MHETase, we hypothesized that the former could confer MHET activity, but abolish PET hydrolytic potential, whereas the latter was expected to have the opposite effect. The lidded PETase was created by replacing the seven-residue loop of PETase (Trp185:Phe191, PETase numbering) with Gly251:Thr472 from MHETase. In control experiments, wild-type PETase exhibited no detectable activity on MHET, and the lidded PETase is not able to degrade amorphous PET film. However, meager activity of lidded PETase was detected on MHET (k_{cat}=0.11±0.02 s^{-1}) (Figure 5.3F). The lidless MHETase was created by replacing the MHETase lid (Gly251:Thr472) with the seven-residue loop of PETase (Trp185:Phe191, PETase numbering). This construction results in an exposed MHETase active site, possibly allowing for acquired PET hydrolytic activity. The resulting enzyme has a k_{cat} value on MHET of 0.05 ± 0.03 s^{-1}, 1,000-fold lower than the rate for wild-type MHETase, demonstrating that the lid domain is crucial for MHET hydrolytic activity (Figure 5.3F). The lidless MHETase enzyme was also unable to degrade amorphous PET film over 96 h, despite the more accessible active site.

Similarly, variants of lidless MHETase were generated to remove the active site disulfide and replace the two sites with tryptophan and serine (lidless MHETase C224W/C529S, see Figure A3.1B of Appendix A3) to reconstitute the PETase active site, or with histidine and phenylalanine (lidless MHETase C224H/C529F), matching the active site of the double-mutant PETase variant previously shown to exhibit improved PET

hydrolytic activity on crystalline PET.[81] The lidless MHETase C224W/C529S mutant, which reconstitutes the wild-type active site motif of PETase, displays the same turnover rate (within error) as the lidded PETase mutant on MHET (k_{cat}=0.10±0.06 s^{-1}), while the lidless MHETase C224H/C529F mutant is even less active on MHET (k_{cat}=0.06±0.03 s^{-1}) (Figure 5.3F*)*. We also generated a PETase variant to recreate the active site disulfide found in MHETase (PETase W159C/S238C). The PETase mutant exhibited very low MHET hydrolytic activity (k_{cat}=0.03±0.3 s^{-1}), and similarly had no activity on BHET or amorphous PET film.

To delineate the effects of engineering the lid and removing the active site disulfide bond, we also generated three MHETase mutants altering only the active site disulfide motif (C224A/C529A, C224W/C529S, and C224H/C529F). We hypothesized that removal of this disulfide bond may diminish the thermal stability of MHETase. However, each of these variants either expressed in inclusion bodies or did not express at all. Attempts were also made to introduce disulfide motifs into MHETase that are found in PETase (G489C/S530C) or in *Ao*FaeB. To recapitulate the *Ao*FaeB disulfide, the mutations include both a point mutation (S136C) as well as the insertion of a 15-residue loop from *Ao*FaeB that harbors the partnering cysteine residue. As with the active site disulfide mutants, these mutants either expressed in inclusion bodies or did not express at all. A variant was also created that included both the PETase-like disulfide (G489C/S530C) and the *Ao*FaeB modification (S136C with 15-residue loop from *Ao*FaeB). This last variant, with seven total disulfides, was successfully expressed and had very low activity on MHET (k_{cat}=0.16±0.14 s^{-1}) (Figure 5.3F).

5.4.6 MHETase is catalytically inactive on MHE-isophthalate and MHE-furanoate.

We evaluated the substrate specificity of MHETase using the mono-hydroxyethyl monomer unit of two additional compounds. Specifically, assays were performed with mono-(2-hydroxyethyl)-isophthalate (MHEI) (Figure A3.12 of Appendix A3) and mono-(2-hydroxyethyl)-furanoate (MHEF) (Figure A3.13 of Appendix A3). Isophthalate is a common co-monomer in industrial PET formulations used to modify crystallinity, such that MHEI could be released from polyester depolymerization. PETase has been demonstrated to deconstruct other aromatic polyesters,[81] including polyethylene furanoate (PEF), yielding MHEF as a product of the enzymatic hydrolysis reaction. Over the course of 24 h at 30°C, no MHETase activity was detected for either substrate using substrate concentrations from 25-250 μM, in contrast with complete hydrolysis of MHET (Figure A3.14 of Appendix A3) in the same time using identical reaction conditions.

To explain the inability of MHETase to act on MHEI and MHEF, we conducted flexible ligand/flexible receptor docking simulations and predicted ten binding orientations for each molecule (MHET, MHEI, and MHEF) in MHETase. These docking simulations indicate that MHET binds to MHETase with a binding free energy of -7.13 kcal/mol and in a catalytically primed configuration. This binding mode features the carbonyl C of MHET within 3.2 Å of Ser225-O, which itself is within 2.90 Å of His528-N(e) and His528-N(d) is 3.93 Å from Asp492-O (Figure A3.15 of Appendix A3). For MHEI and MHEF, no binding modes were predicted that exhibit similarly favorable binding free energies, feature the MHET carbonyl C within range for attack by Ser225, and stabilize the carbonyl of the ester in the oxyanion hole, suggesting that MHETase will not readily act on these molecules.

5.4.7 PETase and MHETase act synergistically during PET depolymerization.

While MHET is susceptible to hydrolysis by a number of PET-degrading cutinases, *I. sakaiensis* requires the action of two enzymes for PET degradation to liberate TPA and EG.[72] Given the turnover rates for MHETase reported here, depolymerization by PETase is likely the rate limiting step when the enzymes are employed together. To investigate the action of the two-enzyme system, we thus measured the extent of hydrolysis of a commercial amorphous PET substrate over 96 h at 30°C using PETase and MHETase at varying concentrations (Figure 5.4A, Table A3.4 of Appendix A3). As expected, MHETase alone has no activity on PET film. Over the range of enzyme loadings tested (0-2.0 mg enzyme/g PET), degradation by PETase alone, as determined by concentration of product released (the sum of BHET, MHET, and TPA), scales with enzyme loading. Upon addition of MHETase in the reaction, at any loading tested (0.1 – 1.0 mg MHETase/g PET), product release still scales with PETase loading, but at a markedly higher level than with PETase alone (Figure 5.4A). The overall trend of degradation within the range of enzyme loadings tested, which shows increasing levels of constituent monomers released as concentration of both enzymes increases, is indicative that these reactions are enzyme-limited under these conditions, rather than substrate-limited. The synergy study does not strongly indicate that any particular ratio of PETase to MHETase results in optimal degradation over the enzyme loadings tested, but rather that degradation scales with PETase loading and the presence of MHETase, even at low concentrations relative to PETase, improves total degradation.

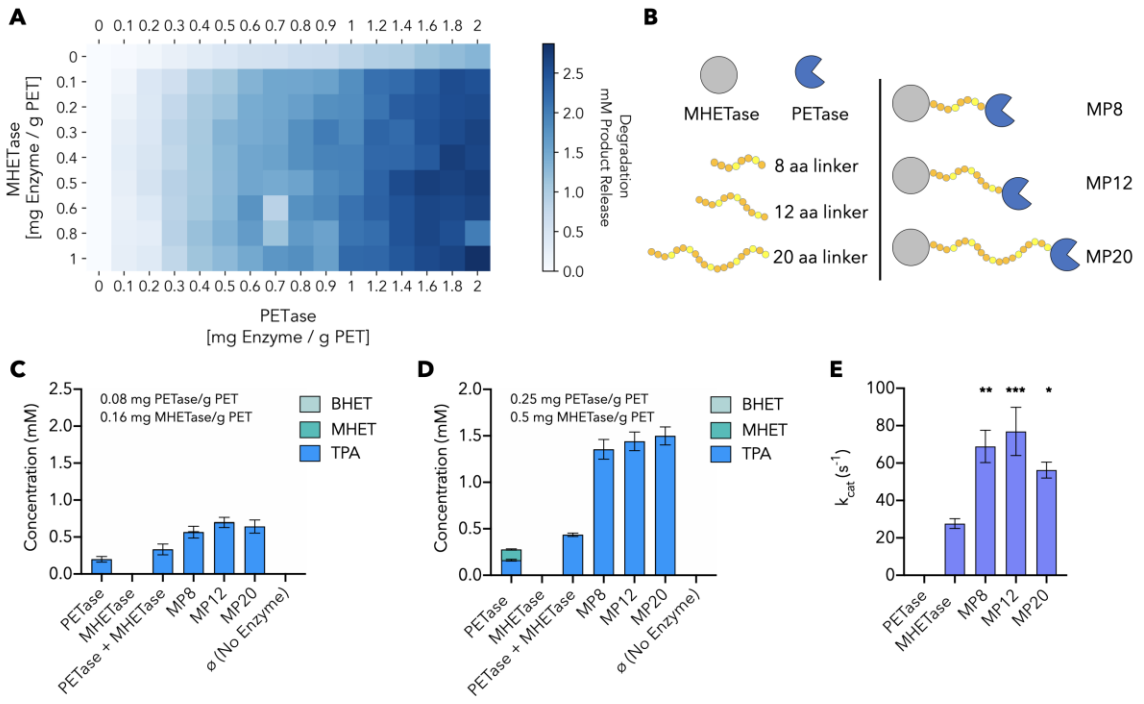


Figure 5.4 PETase-MHETase synergy and chimeric enzymes. (A) Heatmap of synergistic degradation by PETase and MHETase on amorphous PET film over 96 hours at 30°C. Total product release in mM (sum of BHET, MHET, and TPA), x-axis: PETase loading (mg/g PET), y-axis: MHETase loading (mg/g PET). (B) Illustrations of three chimeric enzymes. Linkers composed of glycine (orange) and serine (yellow) residues connecting the C-terminus of MHETase to the N-terminus of PETase. (C, D) Comparison of depolymerization performance of PETase alone, MHETase alone, PETase and MHETase at equimolar loading, and the three chimeric enzymes on amorphous PET film after 96 h at 30°C. Product release in mM resulting from hydrolysis by (C) 0.08 mg PETase/g PET or 0.16 mg MHETase/g PET and (D) 0.25 mg PETase/g PET or 0.5 mg MHETase/g PET. All comparisons are statistically significant with *p*-values ≤ 0.0001 based on 2way ANOVA analysis and Tukey's multiple comparisons test. (E) MHET turnover rate by each chimeric enzyme compared to MHETase alone using 250 µM MHET and 5 nM enzyme.

Asterisks indicate statistically significant comparisons between MHETase and each chimera enzyme with *p*-values ≤ 0.01 (*), 0.001 (**), and 0.0005 (***).

5.4.8 Chimeric proteins of MHETase and PETase improves PET degradation and MHET hydrolysis rates.

In light of the highly synergistic relationship between PETase and MHETase on amorphous PET, where increasing loading of each enzyme results in more constituent monomer release, we next examined how proximity of the two enzymes influences hydrolytic activity. Chimeric proteins covalently linking the C-terminus of MHETase to the N-terminus of PETase using flexible glycine-serine linkers of 8, 12, and 20 total glycine and serine residues were generated and assayed for degradation of amorphous PET (Figure 5.4B). Varying linker lengths were explored to understand the effect of increased mobility between the two domains.[345] Furthermore, two enzyme loadings were compared – the lower loading corresponding to approximately 0.08 mg PETase/g PET and 0.16 mg MHETase/g PET, and the higher enzyme loading corresponding to 0.25 mg PETase/g PET and 0.5 mg MHETase/g PET (Figure 5.4C-D). At both loadings, when comparing the extent of degradation achieved by PETase alone, MHETase alone, and an equimolar mix of PETase and MHETase, the chimeric proteins outperform PETase, as well as the mixed reaction containing both PETase and MHETase unlinked in solution. Furthermore, the chimeras demonstrate a higher catalytic activity on MHET (Figure 5.4E). Chimeric constructs linking the C-terminus of PETase to the N-terminus of MHETase did not successfully express protein (Figure 5.4B). SEM analysis of digested amorphous PET film confirms degradation (Figure A3.16 of Appendix A3).

5.5 Discussion

The ability to degrade polymers to their monomeric units is important for subsequent reuse in new products, which is a critical technical advance needed to enable a global circular materials economy. In biological systems, complete depolymerization to monomers can be necessary for microbial uptake and growth, as in *I. sakaiensis* wherein MHETase is the enzymatic partner to PETase, together allowing for the complete degradation of PET to TPA and EG for catabolism.[72] Prior studies presenting MHETase crystal structures focused upon understanding and tuning substrate specificity, particularly the rational engineering of MHETase to impart BHET hydrolysis activity.[85, 86] Drawing inspiration from our structural analyses, this complementary study offers further insights into the two-enzyme PETase-MHETase system.

The recent structural report from Palm *et al.* highlighted several important amino acid contributions to substrate specificity in MHETase,[85] specifically focusing on active site residues. Of note, they pointed out the importance of Phe415 for substrate binding via an "induced fit" mechanism and highlighted Arg411 with respect to hydrogen bonding of the MHET carboxylate group, both of which are proposed to be drivers of substrate specificity. In addition, beyond engineering a starting point for BHET activity in MHETase for further optimization, the potential for MHEF turnover was suggested, given the proposed utility of PEF as a bio-based PET replacement.[346] In our previous work,[81] we demonstrated that PETase effectively depolymerizes PEF, but the results here do not indicate the same for MHETase on MHEF, and docking simulations agree with the observed patterns in MHETase selectivity. Despite success with predicting a low energy catalytically competent binding mode for MHET to MHETase, we were only able to

predict one binding mode of MHEF to MHETase with the MHEF nucleophilic carbonyl in the oxyanion hole, but in this pose, the MHEF carboxylate moiety is not in range to interact with Arg411, suggesting that further active site engineering will be necessary to enable MHEF turnover. Similarly, only one binding mode for MHEI was predicted wherein the catalytic triad was oriented for catalysis, but, akin to MHEF, the non-linearity of the molecule prevents simultaneous interaction with the oxyanion hole and R411.

The enzyme kinetics studies presented here reveal a substantial reduction in activity for the S131G, E226T, and F495I MHETase mutants, indicating that these positions play important roles in substrate specificity and catalytic efficiency. A previous study also demonstrated greatly reduced hydrolytic activity by a F415S variant.[86] Additionally, two homologs identified via bioinformatics analysis from *Comamonas thiooxydans* and *Hydrogenophaga* sp. PML113 exhibit extremely similar active site environments (Figure 5.3D), with the only exceptions being variations at positions 131 and 415 (MHETase numbering), and these homologs display reaction efficiencies (k_{cat}/K_m) reduced by an order of magnitude (Table A3.3 of Appendix A3). Furthermore, as the amino acids at these positions in wild-type MHETase are less common in tannase family sequences (Figure A3.9 of Appendix A3), and mutation to the more common amino acids led to a reduction in activity, this suggests that these two sequence positions were specifically evolved in MHETase to accommodate MHET.

Two-enzyme systems for complete PET degradation have been examined previously, either derived from a single microorganism (e.g. *Thermobifida fusca*),[347] or screened from multiple sources for optimal activity.[320, 348] The enzyme synergy results for the *I. sakaiensis* PETase -MHETase system on amorphous PET display a clear performance

improvement when MHETase is included in the reaction. Namely, overall degradation scales with PETase loading within the tested range (0 – 2.0 mg PETase/g PET), but the inclusion of MHETase in the degradation reaction markedly improves depolymerization and this synergistic enhancement also scales with MHETase loading.

The presence of confirmed MHETase homologs in *C. thiooxydans* and *Hydrogenophaga* sp. PML113 suggests that these bacteria may harbor abilities for TPA catabolism (Figure A3.17 of Appendix A3). Bioinformatics analysis was thus conducted to query the genomes of the strains compared to known TPA catabolic genes from *I. sakaiensis,*[72] *Comamonas* sp. E6,[349, 350] *Delftia tsuruhatensis,*[351] *Paraburkholderia xenovorans,*[352] *Rhodococcus jostii* RHA1,[353] and *Rhodococcus* sp. DK17,[354] including putative PETases, terephthalate transporter genes, two-component terephthalate dioxygenases, the 1,2-dihydroxy-3,5-cyclohexadiene-1,4-dicarboxylate dehydrogenase, and the three types of protocatechuate (PCA) dioxygenases (PCA-2,3, PCA-3,4, and PCA-4,5-dioxygenases) (Table A3.5 of Appendix A3). This analysis revealed that neither *C. thiooxydans* nor *Hydrogenophaga* sp. PML113 harbor putative PETase genes. Interestingly, both strains exhibit genes encoding for TPA catabolic enzymes and transporters highly homologous to those of *I. sakaiensis, Comamonas* sp. E6, and *Delftia tsuruhatensis* (in all cases above 60% identity (Figures A3.18 and A19 of Appendix A3), suggesting that they are highly likely able to turnover TPA to PCA, a common central intermediate in aerobic aromatic catabolic pathways.[355] Each strain also contains annotated PCA-4,5-dioxygenases (Table A3.5 of Appendix A3). Further experimental work will be required to understand if either of these bacteria exhibit the ability to depolymerize PET, perhaps through another type of mechanism than via ester hydrolases, or perhaps like

Comamonas sp. E6, they are primarily able to consume soluble, xenobiotic intermediates. Perhaps these strains could serve as useful sources of TPA catabolic genes for synthetic biology efforts associated with biological plastics recycling and upcycling.[356]

The enzymatic deconstruction of recalcitrant natural polymers such as cellulose, hemicellulose, and chitin is accomplished in nature by the action of cocktails of synergistic enzymes secreted from microbes.[12, 357] For example, as observed in fungal cellulase systems for cellulose depolymerization, these cocktails typically contain a subset of enzymes to act directly on solid polymeric substrates via interfacial enzyme mechanisms, and complementary enzymes (e.g., β-glucosidases) that further process solubilized intermediates to monomeric constituents (e.g., cellobiose hydrolysis to glucose). Given that natural microbial systems evolved over millions of years to optimally degrade recalcitrant polymers, perhaps it is thus not surprising, in hindsight, that a soil bacterium such as *I. sakaiensis* evolved the ability to utilize a crystalline polyester substrate with, to our collective knowledge, a two-enzyme system.[72, 78] Extending the analogy of cellulase enzymes and plant cell wall deconstruction for breaking down diverse polysaccharides simultaneously, the concept of deconstructing synthetic polymers in the form of mixed plastics waste with advanced enzyme cocktails is an exciting research direction beyond PET to other polyesters, natural fibers (e.g., cellulose from cotton, proteins from wool),[358] polyamides, polyurethanes,[359] and other polymers susceptible to enzymatic depolymerization. Going forward, the design of multi-enzyme systems for depolymerization of mixed polymer wastes is a promising and fruitful area for continued investigation.

5.6 Methods

5.6.1 Plasmid construction.

pET21b(+)-based expression plasmids for *I. sakaiensis* genes, homologous genes, and mutants were generated as further described in Appendix A3.

5.6.2 Protein expression and purification.

E. coli-based protein expression and chromatographic purification is described in Appendix A3.

5.6.3 Crystallization and structure determination.

MHETase was crystallized in four conditions including a seleno-methionine-labeled version for single-wavelength anomalous diffraction phasing. All X-ray data collections were performed at Beamline I03 at the Diamond Light Source. Detailed methods and statistics are provided in Appendix A3.

5.6.4 Molecular simulations.

MD simulations were performed for solvated MHETase both in the free state and with MHET bound at the active site. All systems were built in CHARMM,[329] and simulations utilized the CHARMM forcefield.[331] Classical MD simulations were run with NAMD;[330] QM/MM simulations, including two-dimensional umbrella sampling free energy calculations, were run in Amber.[332, 360] Additional simulation details are in Appendix A3.

5.6.5 Bioinformatics.

A total of 6,671 tannase family sequences were retrieved via PSI-BLAST against the NCBI non-redundant database.[340] Phylogenetic analyses were conducted with MEGA7.[361] Additional details are in Appendix A3.

5.6.6 MHETase kinetics and turnover experiments.

MHETase and mutant enzymes were incubated with MHET, MHEI, or MHEF and reactions quenched with methanol and a heat treatment at 85˚C for 10 min. Hydrolysis extent was measured by HPLC as described in Appendix A3.

5.6.7 Molecular docking.

MHET, MHEI, and MHEF docking into MHETase were modeled and prepared using tools in Schrödinger. Substrate docking simulations were conducted using Induced Fit Docking simulations in Schrodinger as described in Appendix A3.

5.6.8 Ligand synthesis.

MHET, MHEI, and MHEF were prepared via the coupling and subsequent deprotection of a mono-tBoc-protected ethylene glycol with the respective acyl chlorides as further described in Appendix A3.

5.6.9 MHETase synergy with PETase.

The effect of MHETase loading and PETase loading on amorphous PET film after 96 hours was measured as total product release (MHET, BHET and TPA) via HPLC, as described in Appendix A3.

5.6.10 MHETase-PETase chimeras.

Chimeric constructs covalently linking MHETase to PETase were generated and incubated with either MHET or amorphous PET film as described in Appendix A3.

CHAPTER 6. Predicting Protein Thermostability with Machine Learning.

This chapter presents a machine learning method, ThermoProt, for discriminating psychrophilic, mesophilic, thermophilic, and hyperthermophilic proteins. The author of this dissertation performed all computational experiments and analyses.

6.1 Abstract

In silico prediction of protein thermostability is of vital relevance in biotechnology design and implementation, as exhaustive experimental determination of thermostability is not feasible for the extremely large number of proteins in publicly available databases. *In silico* prediction methods, conversely, can be applied to large protein sets to select a smaller library with desired thermal properties. Hence, computational tools that enable high-throughput prediction of thermostability would immensely facilitate the process of protein engineering. In this study, we present a machine-learning method, ThermoProt, that uses a support vector machine (SVM) algorithm for predicting the thermostability of proteins. Predictive accuracies of 74.0%, 85.5%, 83.3%, and 86.6% were obtained in discriminating psychrophilic from mesophilic, mesophilic from thermophilic, thermophilic from hyperthermophilic, and mesophilic from thermophilic and hyperthermophilic proteins, respectively. Compared to other previous methods, ThermoProt shows competitive performance. We also conducted statistical studies of amino acid correlations to investigate relationships between the mutual constraints of amino acids and thermostability. We determined that the pairwise correlations of 63 amino acid pairs were significantly different ($p < 0.01$) between psychrophilic, mesophilic, thermophilic, and hyperthermophilic

proteins. ThermoProt is available as a free Python package at https://github.com/jafetgado/ThermoProt and can be downloaded from the Python Package Index at https://pypi.org/project/ThermoProt.

6.2 Introduction

Proteins with high thermal stability are especially desirable for industrial applications in biocatalysis. Enzymes with higher thermal stability allow biochemical processes to be conducted at elevated temperatures, leading to faster reaction rates. As a result, the discovery and engineering of proteins with robust, high-temperature behavior is a major research area. For example, thermostable amylases have been successfully employed in the production of glucose from starch at temperatures as high as 100°C.[362-364]

Based on their optimum growth temperatures, organisms are classified as psychrophilic, mesophilic, thermophilic, or hyperthermophilic, which roughly correspond to the temperatures 20°C or less, 20°C to 45°C, 45°C to 80°C, and 80 ˚C or more, respectively.[365, 366] For proteins isolated from natural systems, there is a fairly linear relationship between the protein's thermal stability and the environmental temperature of the native organism. Gromiha *et al*. initially observed the linear relationship, $T_m = 24.4°C + 0.93T_{env}$ (correlation coefficient = 0.91), between the melting temperature (T_m) of 56 proteins and the organisms' average environmental temperature (T_{env}).[367] From an analysis of 127 proteins, Dehouck *et al*. (2008) observed a similar linear relationship but with a much lower correlation of 0.59.[368] Since T_m data for proteins is scarce, the significant linear relationship between T_m and T_{env} can be practically exploited in a computational framework to identify proteins that are very likely to be thermostable.

The mechanistic strategies employed by organisms in their proteomes to adapt to extreme temperatures have been long examined. From these studies, many features have been suggested as hallmarks of protein thermostability, such as an increase in helical content and polar surface composition.[369, 370] However, different protein families appear to adopt unique strategies, such that a singular paradigm may not be universally applicable to all protein families.[371] Kumar *et al*. studied several factors thought to contribute to thermostability in mesophilic and thermophilic representatives of 18 protein families, including increased proline occurrence in loops, shortening and deletion of loops, and hydrogen bonding.[372] They found that, although there were strong trends between some of these factors and thermostability, no single factor was consistent in all 18 families. Therefore, regardless of overall trends, separate studies may be required for specific protein families to identify unique features adopted by the thermostable members of the family.

While there appears to be no one universal rule for delineating thermal stability consistently across all protein families, research indicates there are many general relationships between protein features and thermostability. An increase in salt bridges, ionic interactions, and hydrogen bonds have been commonly observed to favor thermostability, and thermophilic proteins improve these interactions by utilizing a higher proportion of charged residues at the expense of uncharged polar residues.[371-374]. Similarly, thermophilic proteins employ a higher concentration of hydrophobic residues to increase hydrophobic interactions and rigidity.[375, 376] Furthermore, an increase in Gibbs free energy change of hydration ($-G_{hN}$), shape factor(s), disulfide bridges, cation-π interactions, and aromatic clusters have been correlated with protein thermostability.[367, 376, 377] From combinatorial studies, Farias *et al*. discovered that the composition ratio, (Glu + Lys)/(Gln

+ His), over the proteomes of 28 organisms strongly correlated with the growth temperatures.[378] Similarly, Zeldovich *et al*. determined that the sum of the compositions of Ile, Val, Tyr, Trp, Arg, Glu, and Leu in proteomes had a near-perfect correlation (0.93) with the growth temperatures of 204 organisms.[379]

In recent years, researchers have sought to develop statistical tools to predict protein thermostability. Several machine-learning methods have been employed, including decision trees, random forests, support vector machines (SVMs), k-nearest neighbor (KNN), and neural networks.[380-390] Validation tests of these methods have shown percentage accuracies ranging from the low 70s to the mid-90s. However, direct comparison of the predictive performance of these methods can be misleading since widely differing datasets and evaluation procedures were applied. In each of these studies, the amino acid composition has been the most powerful predictors of thermostability. Algorithms built on dipeptide composition alone have underperformed methods based on the amino acid composition.[385] Other features such as evolutionary information, secondary and tertiary structure characteristics, and physiochemical properties have been applied for machine learning in addition to amino acid composition. In isolation, these alternate features did not increase prediction accuracy over the amino acid composition feature but, when applied in combination with amino acid composition, have led to an modest improvement in performance.[384, 385, 390] Since protein structure is a deterministic function of the amino acid sequence, it is no surprise that robust machine-learning methods have been reasonably successful in predicting thermostability using only sequence information.

In this work, we apply machine learning to predict protein thermostability using a unique combination of carefully selected features on a larger and more diverse dataset than

has been previously used. We combine top features that have previously been determined to correlate with thermostability, with top features from a feature selection technique. Our algorithm is made available to the public as a Python module (ThermoProt) via the permanent, open-access database GitHub and on the Python package repository (PyPI). We also investigate statistically significant correlations between amino acid composition and thermostability on our dataset to provide insight into the constraints of amino acid occurrence in thermostable proteins.

6.3 Materials and methods

6.3.1 Sequence dataset

We retrieved sequence data for three psychrophilic, three mesophilic, six thermophilic, and eight hyperthermophilic organisms from the National Center for Biotechnology Information (NCBI) database, totaling 234,171 sequences (Table 6.1). A 40% sequence-identity threshold was applied using the CD-HIT algorithm.[391] 40,000 sequences were selected from the CD-HIT output such that there were 10,000 sequences in each class (P: psychrophilic, M: mesophilic, T: thermophilic, H: hyperthermophilic). 8,000 of these sequences (2,000 in each class) were set aside for hyperparameter optimization and feature selection, while the remaining 32,000 sequences were used for training, validation, and analysis. Protein sequences of *Rhodonelium psychrophilium* (P), *Methylobacillus flagellates* (M), *Ardenticatena maritima* (T), and *Thermotoga petrophila* (H) were retrieved from NCBI to constitute a separate dataset of 22,299 proteins for an independent test of the final algorithm (Table 6.2).

Table 6.1 Organisms and protein sequences for feature selection and validation dataset. Optimum temperatures were retrieved from NCBI Bioproject database (https://www.ncbi.nlm.nih.gov/bioproject/) and BacDive (https://bacdive.dsmz.de/)

S/No	Organism	Group	Growth/Optimum Temp (˚C)		Number of Proteins	Class
			NCBI	BacDive		
1.	*Psychroflexus torquis*	Bacteria	0-15	4	9,953	Psychrophilic
2.	*Moritella sp.*	Bacteria	5-8	5	31,433	
3.	*Colwellia psychrerythraea*	Bacteria	8	10	31,845	
4.	*Vibrio mediterranei*	Bacteria	26	25-28	48,521	Mesophilic
5.	*Parvimonas micra*	Bacteria	37	37	12,460	
6.	*Aeromonas enteropelogenes*	Bacteria	36	30	27,934	
7.	*Thermogemmatispora onikobensis*	Bacteria	60-65	60-65	4,255	Thermophilic
8.	*Thermovenabulum gondwanense*	Bacteria	65	65	4,424	
9.	*Acidianus brierleyi*	Archaea	70	70	10,479	
10.	*Metallosphaera sedula*	Archaea	70	65	18,352	
11.	*Thermomicrobium roseum*	Bacteria	70	70	5,641	
12.	*Thermobifida fusca*	Bacteria	50-55	45-60	19,415	
13.	*Methanocaldococcus vulcanius*	Archaea	80	80	3,446	Hyperthermophilic
14.	*Thermococcus sp.*	Archaea	85	80	39,447	
15.	*Vulcanisaeta distributa*	Archaea	85-90	90	4,921	
16.	*Geoglobus ahangari*	Archaea	88	85	3,958	
17.	*Thermococcus guaymasensis*	Archaea	88	88	4,121	
18.	*Aeropyrum pernix*	Archaea	90-95	90-95	18,861	
19.	*Pyrococcus kukulkanii*	Archaea	105	105	4,061	
20.	*Pyrolobus fumarii*	Archaea	106	103	3,875	

Table 6.2 Organisms and protein sequences for separate testing set.

S/No	Organism	Group	Growth/Optimum Temp (°C)		Number of Proteins	Class
			NCBI	BacDive		
1	*Rhodonellum psychrophilum*	Bacteria	5	5 - 28	5,035	Psychrophilic
2	*Methylobacillus flagellatus*	Bacteria	30 - 42	30	5,743	Mesophilic
3	*Ardenticatena maritima*	Bacteria	60	62 - 65	8,881	Thermophilic
4	*Thermotoga petrophila*	Bacteria	80	80	2,640	Hyperthermophilic

6.3.2 Feature selection

Since our goal was to develop efficient and versatile predictive algorithms, we focused on features derived from the amino acid sequence alone and avoided structure-based features. This ensures that our algorithms can be readily applied in the absence of a crystal structure. Moreover, in several previous works, the addition of structure-based features did not lead to very significant improvements in performance.[384, 392, 393] We applied three categories of features:

1. Amino acid composition (AAC) features: the fractional amounts of 20 canonical amino acids in the proteins.

2. g-gap dipeptide composition (DPC) features: the relative amounts of $a(x)_g b$, where a and b are specific amino acids and $(x)_g$ represents g amino acids of any type, sandwiched between a and b.[394] In this work, we tested 1,200 g-gap dipeptides (i.e., g = 0, 1, and 2).

3. Residue type and physiochemical (RTP) features: we selected 20 residue-type and physicochemical features that have been previously determined to significantly correlate with thermal stability, namely, the composition of acidic, basic, non-polar, acyclic, aliphatic, aromatic, charged, and Glu + Phe + Met + Arg residues; the ratio of basic to

acidic, non-polar to polar, acyclic to cyclic, and charged to non-charged residues;[383] the composition of tiny (Ala, Gly, Pro, Ser) and small (Thr, Asp) residues, the average maximum solvent accessible area (ASA),[392] the ratio of (Glu + Lys) to (Gln + His),[378] charged vs. polar composition,[395] IVYWREL (Ile, Val, Tyr, Trp, Arg, Glu, Leu) composition,[379] molecular weight, and heat capacity.[133]

6.3.3 Learning and evaluation

We applied six machine-learning methods: random forests, logistic regression, Gaussian naïve Bayes, K-nearest neighbor (KNN), support vector machines with linear kernel (SVM), and support vector machines with radial basis function kernel (RBF SVM). These methods were implemented using the Scikit-learn Python package.[396] A multi-label random forest classifier was trained on the separate optimization dataset (8,000 sequences) using only the 1,200 g-gap DPC features. The top 10 features with the highest Gini feature importances were selected,[136] and the rest were discarded. Optimum hyperparameters for the KNN, RBF SVM, and random forest classifiers were determined by 5-fold cross validation on the dataset of 8,000 sequences. For each of the six machine-learning methods, four types of binary classifiers were trained and tested: psychrophilic vs. mesophilic (PM), mesophilic vs. thermophilic (MT), thermophilic vs. hyperthermophilic (TH), and mesophilic vs. thermophilic and hyperthermophilic (MTH). Five-Fold cross validation on the dataset of 32,000 proteins was used to evaluate the performance of the classifiers.The predictive performance of the classifiers was measured using accuracy, sensitivity, specificity, and Matthew's correlation coefficient (MCC), as defined by equations 2.6, 2.7, 2.8, and 2.13 respectively.

Table 6.3 Amino-acid sequence features used in this study and the Spearman's correlation between each feature and the thermostability class (P=1, M=2, T=3, H=4).

S/No	Features	Short Name	Description	Correlation Coefficient (r)
1	A composition	A comp	n_A/n_{total}	0.013
2	C composition	C comp	n_C/n_{total}	-0.081
3	D composition	D comp	n_D/n_{total}	-0.158
4	E composition	E comp	n_E/n_{total}	0.216
5	F composition	F comp	n_F/n_{total}	-0.137
6	G composition	G comp	n_G/n_{total}	0.201
7	H composition	H comp	n_H/n_{total}	-0.162
8	I composition	I comp	n_I/n_{total}	0.010
9	K composition	K comp	n_K/n_{total}	-0.085
10	L composition	L comp	n_L/n_{total}	0.071
11	M composition	M comp	n_M/n_{total}	-0.039
12	N composition	N comp	n_N/n_{total}	-0.318
13	P composition	P comp	n_P/n_{total}	0.196
14	Q composition	Q comp	n_Q/n_{total}	-0.427
15	R composition	R comp	n_R/n_{total}	0.358
16	S composition	S comp	n_S/n_{total}	-0.258
17	T composition	T comp	n_T/n_{total}	-0.182
18	V composition	V comp	n_V/n_{total}	0.316
19	W composition	W comp	n_W/n_{total}	0.024
20	Y composition	Y comp	n_Y/n_{total}	0.088
21	AA 0-gap dipeptide composition	AA 0-gap	$n_{AA}/(n_{total}-1)$	-0.033
22	RE 0-gap dipeptide composition	RE 0-gap	$n_{RE}/(n_{total}-1)$	0.274
23	RR 0-gap dipeptide composition	RR 0-gap	$n_{RR}/(n_{total}-1)$	0.272
24	EQ 0-gap dipeptide composition	EQ 0-gap	$n_{EQ}/(n_{total}-1)$	-0.230
25	QA 0-gap dipeptide composition	QA 0-gap	$n_{QA}/(n_{total}-1)$	-0.198

Table 6.3 (continued)

S/No	Features	Short Name	Description	Correlation Coefficient (r)
26	KQ 0-gap dipeptide composition	KQ 0-gap	$n_{KQ}/(n_{total}-1)$	-0.240
27	R*R 1-gap dipeptide composition	R*R 1-gap	$n_{R*R}/(n_{total}-2)$	0.187
28	A**R 2-gap dipeptide composition	A**R 2-gap	$n_{A**R}/(n_{total}-3)$	0.095
29	L**Q 2-gap dipeptide composition	L**Q 2-gap	$n_{L**Q}/(n_{total}-3)$	-0.302
30	R**R 2-gap dipeptide composition	R**R 2-gap	$n_{R**R}/(n_{total}-3)$	0.259
31	Acidic residue composition	Acidic	$\sum n_x/n_{total}$ for x in [D,E]	0.081
32	Basic residue composition	Basic	$\sum n_x/n_{total}$ for x in [K,R,H]	0.141
33	Non-polar residue composition	Non-polar	$\sum n_x/n_{total}$ for x in [A,G,I,L,M,F,P,W,V]	0.231
34	Cyclic residue composition	Cyclic	$\sum n_x/n_{total}$ for x in [F,Y,W,P,H]	0.021
35	Aliphatic residue composition	Aliphatic	$\sum n_x/n_{total}$ for x in [A,G,I,L,V]	0.248
36	Aromatic residue composition	Aromatic	$\sum n_x/n_{total}$ for x in [H,F,W,Y]	-0.094
37	Charged residue composition	Charged	$\sum n_x/n_{total}$ for x in [D,E,K,R,H]	0.131
38	Basic/acidic ratio	Basic/acidic	$Basic/Acidic$	0.019
39	Non-polar/polar ratio	Non-pol/pol	$Non\text{–}polar/(1-Non\text{–}pc$	0.192
40	Cyclic/acyclic ratio	Cyc/acyc	$Cyclic/(1-Cyclic)$	0.023
41	Charged/non-charged ratio	Charged/non	$Charged/(1-Charged)$	0.152
42	EFMR composition	EFMR comp	$\sum n_x/n_{total}$ for x in [E, F, M, R]	0.310
43	(E+K)/(Q+H)	(E+K)/(Q+H)	$\frac{n_E+n_K}{n_Q+n_H}$	0.290
44	CvP	CvP	$\sum n_x/n_{total} - \sum n_y/n_{total}$ for x in [D,E,K,R], for y in [N,Q,S,T]	0.396

Table 6.3 (continued)

S/No	Features	Short Name	Description	Correlation Coefficient (r)
45	IVYWREL composition	IVYWREL	$\sum {}^{n_x}/_{n_{total}}$ for x in [I,V,Y,W,R,E,L]	0.525
46	Tiny residues composition	Tiny res	$\sum {}^{n_x}/_{n_{total}}$ for x in [A,G,P,S]	0.062
47	Small residues (TD) composition	Small res	$\sum {}^{n_x}/_{n_{total}}$ for x in [T,D]	-0.246
48	Average maximum ASA	ASA	$\sum ({}^{n_x}/_{n_{total}} \times A_x)$ for x in all 20 amino acids, A_x is the maximum solvent accessible surface area of amino acid, x.	0.023
49	Molecular weight (kDa)	Mol weight	$\sum (n_x \times W_x)$ for x in all 20 amino acids, W_x is the molecular weight of amino acid, x.	-0.063
50	Heat capacity	Heat cap	$\sum ({}^{n_x}/_{n_{total}} \times c_x)$ for x in all 20 amino acids, c_x is heat capacity of amino acid, x.	-0.178

6.4 Results

6.4.1 Evaluation of performance

The top 10 DPC features were selected from the 1,200 0-gap, 1-gap, and 2-gap DPC features using the relative Gini importances of the random forest features.[136] The top 10 DPC features were the AA, RE, RR, EQ, QA, KQ, R*R, A**R, L**Q, and R**R compositions (* represents intervening residues). Hence, a total of 50 features were applied in the algorithms (20 AAC, 20 RTP, 10 selected DPC). From a grid search evaluated with

5-fold cross validation, we determined the optimum hyperparameters for the KNN, random forests, and RBF SVM classifiers on a separate dataset of 8,000 sequences (Table 6.4).

Table 6.4 Optimum hyperparameters for machine learning classifiers determined using the feature selection dataset of 8,000 sequences.

Method	Hyperparameter	PM	MT	TH	MTH
KNN	Number of neighbors (k)	20	20	20	20
Random Forests	Number of trees (n)	200	200	200	200
RBF SVM	Regularization parameter (C)	3.0	3.5	3.0	3.0
RBF SVM	Kernel coefficient (γ)	0.0125	0.0125	0.0100	0.0125

On the validation dataset of 32,000 sequences, 5-fold cross validation was used to evaluate the performance of the classifiers. For each classification instance, there were 8,000 proteins in each class (P, M, T, H). To avoid class imbalance, random undersampling was applied for the MTH datasets, i.e. discrimination was carried out between 8,000 mesophilic (M) proteins and 8,000 thermophilic and hyperthermophilic proteins that were randomly selected from the initial set of 16,000 thermophilic and hyperthermophilic proteins.[397] We found that the RBF SVM outperformed all other methods with overall accuracies of 74.0%, 85.5%, 83.3%, and 86.6% for the PM, MT, TH, and MTH classification, respectively (**Tables 6.5 and 6.6**). Thus, we selected the RBF SVM as the preferred classifier. In previous studies, SVMs have outperformed most classifiers and tend to be the preferred method in predicting thermostability.[381, 382, 385, 386, 390] Moreover, in every method we tested, the MTH classifier showed the highest accuracy, followed by the MT, TH, and PM classifiers, respectively.

Table 6.5 Overall accuracies of classifiers in discriminating psychrophilic from mesophilic proteins (PM), mesophilic from thermophilic proteins (MT), thermophilic from hyperthermophilic proteins (TH), and mesophilic from thermophilic and hyperthermophilic proteins. Accuracies are reported as mean ± the standard deviation over a 5-fold cross validation on the validation set (32,000 proteins).

	PM	MT	TH	MTH
Logistic regression	71.0 ± 0.9	80.5 ± 0.9	76.6 ± 0.2	82.4 ± 0.8
KNN	69.6 ± 0.8	83.3 ± 0.4	81.0 ± 0.9	83.6 ± 0.3
Naïve Bayes	68.0 ± 0.9	73.9 ± 0.7	70.8 ± 0.6	77.0 ± 1.2
Random forests	73.0 ± 0.6	84.5 ± 0.5	82.9 ± 0.5	85.2 ± 0.5
Linear SVM	71.3 ± 1.0	80.5 ± 0.2	76.7 ± 0.4	82.3 ± 0.8
RBF SVM	74.0 ± 0.5	85.5 ± 0.4	83.3 ± 0.6	86.6 ± 0.8

Table 6.6 Validation performance of RBF SVM (ThermoProt) measured over a 5-fold cross validation on the validation datasets of 32,000 proteins.

	PM	MT	TH	MTH
Accuracy	74.0 ± 0.5	85.5 ± 0.4	83.3 ± 0.6	86.6 ± 0.8
Sensitivity	76.2 ± 0.7	86.1 ± 0.5	80.5 ± 0.5	87.0 ± 1.2
Specificity	72.1 ± 0.5	85.0 ± 0.5	86.6 ± 1.5	86.3 ± 0.9
MCC	0.48 ± 0.01	0.71 ± 0.01	0.67 ± 0.01	0.73 ± 0.02

6.4.2 Comparison with other methods

Previous studies have employed machine learning to discriminate proteins according to their thermal stability and have reported classifiers with relatively high predictive accuracies. For instance, Wang *et al.* reported an accuracy of 95.93% for an SVM classifier in discriminating thermophilic from mesophilic proteins;[386] however, the size of the sequence dataset that was used in evaluating the performance of the SVM was

notably small (418 proteins). We compared the performance of our RBF SVM method (ThermoProt) with other methods by training and validating other methods on our dataset of 32,000 proteins (8,000 per class) with 5-fold cross validation. We observed that, compared to the original reported performance, there was a significant drop in the performance of other methods, by as much as 15%, when applied to our larger and more diverse dataset (Table 6.7). These results demonstrate the necessity of using sufficiently large and diverse datasets in machine-learning problems, since the use of small datasets may lead to overfitting and an overestimation of performance. Compared to all the other methods we tested, our method performs best when applied to the larger dataset of 32,000 proteins.

Table 6.7 Comparison of ThermoProt with other methods on the MT, MTH, and PM datasets defined in this study.The MT dataset is comprised of 8,000 mesophilic and 8,000 thermophilic proteins. The MTH dataset is comprised of 8,000 mesophilic, and 8,000 thermophilic and hyperthermophilic proteins (4,000 each). Accuracy was measured over a 5-fold cross validation.

	Method	Dataset Size	Reported accuracy	Accuracy on our datasets		
				MT	MTH	PM
Gromiha and Suresh, 2008	SVM	4,684	89.2	79.9	81.9	-
Wu *et al*., 2009	Decision trees	1,810	85.0	69.3	70.1	-
Lin and Chen, 2011	SVM	1,708	93.3	85.0	85.8	-
Nath *et al.*, 2012	Random forests	12,000	69.3	-	-	72.6
ThermoProt (this study)	SVM	16,000	-	85.5	86.6	74.0

ThermoProt also shows better performance when compared on another dataset. Gromiha and Suresh prepared a dataset of 4,684 mesophilic and thermophilic proteins by applying a 40% sequence identity threshold to the dataset used by Zhang and Fang.[381, 382] This dataset of 4,684 proteins has been used by several researchers in developing and testing algorithms for predicting protein thermostability. Fan *et al*. applied SVMs using 460 features on the Gromiha and Suresh dataset and reported an accuracy of 93.53%.[390] They compared their method with other existing methods and observed that their method outperformed all others, including the methods of Zuo *et al* and Lin and Chen.[385, 387] We applied ThermoProt to the Gromiha and Suresh dataset and compared the performance with the other methods using the data presented by Fan *et al*.[390] On the Gromiha and Suresh dataset, our method outperforms every other method except Fan *et al.* (Table 6.7). Unfortunately, while the Fan *et al*. SVM classifier is reported to have slightly higher accuracy on the Gromiha and Suresh data set than ThermoProt (93.5% vs. 91.9%, respectively), the method of Fan *et al*. is not publicly available. Moreover, ThermoProt will likely be much more computationally efficient to implement, since we use only 50 features compared to 460 features used by Fan *et al.*

Table 6.8 Comparison of methods on Gromiha and Suresh dataset. Performance data are derived from Fan et al, 2016.[390]

	Method	Accuracy	Sensitivity	Specificity
Gromiha and Suresh, 2008	Neural network	89.0	83.3	92.0
Wu *et al*., 2009	Decision tree	83.9	81.5	85.2
Lin and Chen, 2011	SVM	90.8	85.4	93.6
Zuo *et al*., 2013	KNN-ID	91.0	84.3	94.5
Fan *et al.*, 2016	SVM	93.5	89.5	95.6
ThermoProt	SVM	91.9 ± 0.8	89.3 ± 2.4	93.1 ± 3.9

6.4.3 Performance on an independent test set

To further test the performance of our method, we applied the PM, MT, TH, and MTH classifiers which were trained on the validation dataset (32,000 proteins) to a separate test dataset of 22,299 proteins completely withheld from the validation process. The performance of the classifiers on this separate dataset is comparable to the results obtained from the 5-fold cross validation (Table 6.9). For example, in discriminating psychrophilic from mesophilic proteins (PM), ThermoProt shows similar accuracies of 74% and 75% on the validation and independent test datasets, respectively.

Table 6.9 Accuracy of classifiers on separate test set.

Organism	Size	Accuracy			
		PM	MT	TH	MTH
R. psychrophilum (P)	5035	75.0	-	-	-
M. flagellatus (M)	5743	87.1	80.1	-	82.5
A. maritima (T)	8881	-	80.2	85.8	77.2
T. petrophila (H)	2640	-	-	86.1	86.9

To investigate the effects of protein sequence length on the predictive accuracy classifiers, we separated the independent dataset into bins according to the number of residues in the proteins and evaluated the performance in each bin. Our results suggest that the predictive accuracy generally increases as protein size increases but may start to decrease as the sequence length exceeds 800 residues (Figures 6.1 and 6.2). A similar trend was observed by Zuo *et al*.[387]

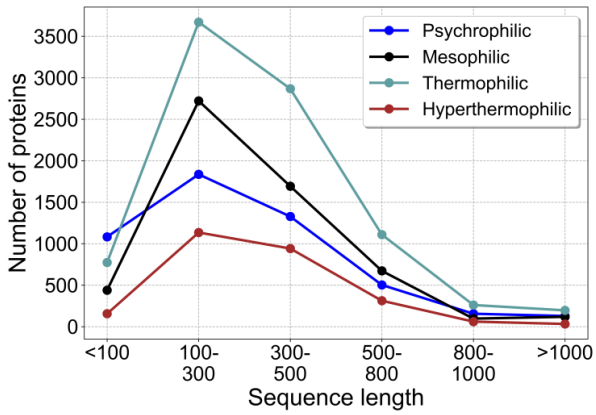


Figure 6.1 Protein size distribution of separate test set of 22,299 proteins. Sequences are of *Rhodonelium psychrophilium* (psychrophilic), *Methylobacillus flagellates* (mesophilic), *Ardenticatena maritima* (thermophilic), and *Thermotoga petrophila* (hyperthermophilic).

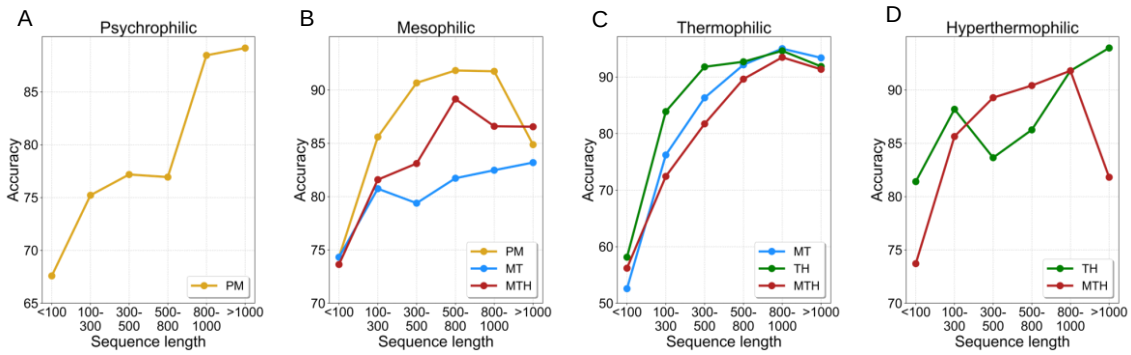


Figure 6.2 Predictive accuracy of RBF SVM classifier (ThermoProt) on independent test set of 22,299 proteins as a function of protein size.Sequences were separated into bins and the classification accuracy was measured over each bin: (A) Psychrophilic proteins (B) Mesophilic proteins (C) Thermophilic proteins (D) Hyperthermophilic proteins.

6.4.4 Amino acid correlations in psychrophilic, mesophilic, thermophilic and hyperthermophilic proteins

Let $r_{x,y}^{S}$ be the Pearson's correlation coefficient between the frequencies of amino acids, x and y, in S, the set of psychrophilic, mesophilic, thermophilic, or hyperthermophilic proteins, and let $\overline{S}$ represent the complement of S. The correlation coefficients, $r_{x,y}^{S}$, for all pairs of amino acids were calculated from the validation dataset (32,000 proteins). The results are shown in Figures 6.3 to 6.6 below. High positive correlation between amino acids, x and y, in the protein set, S, implies that there is a positive constraint in the mutual occurrence of amino acids, x and y, so that an increase in the composition of x is associated with an increase in the composition of y. Similarly, high negative correlation between x and y indicates that x and y are inversely constrained in the set of proteins, such that an increase in x is associated with a decrease in y.

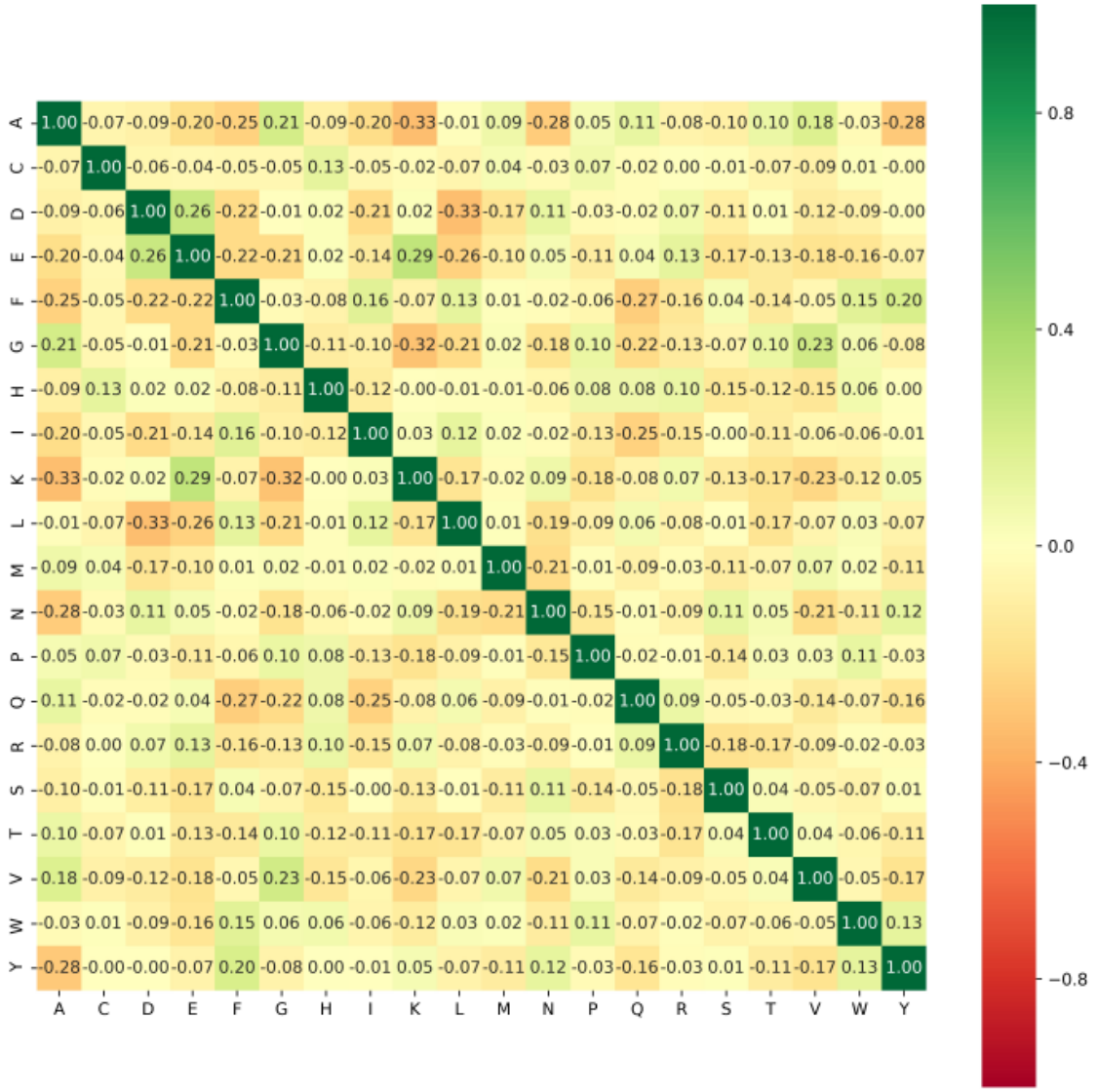


Figure 6.3 Pearson correlation coefficient between amino acid frequencies in psychrophilic proteins.

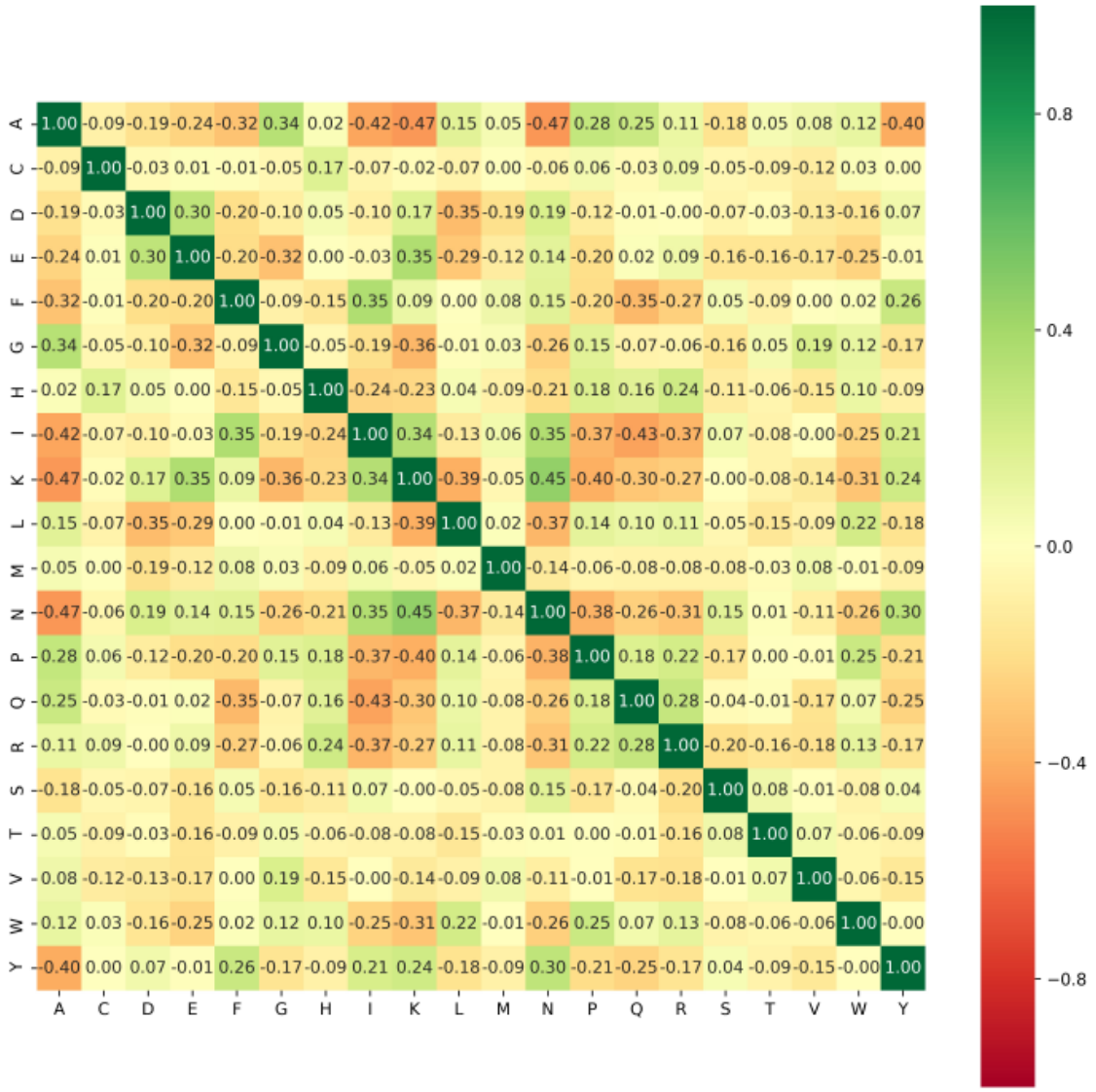


Figure 6.4 Pearson correlation coefficient between amino acid frequencies in mesophilic proteins.

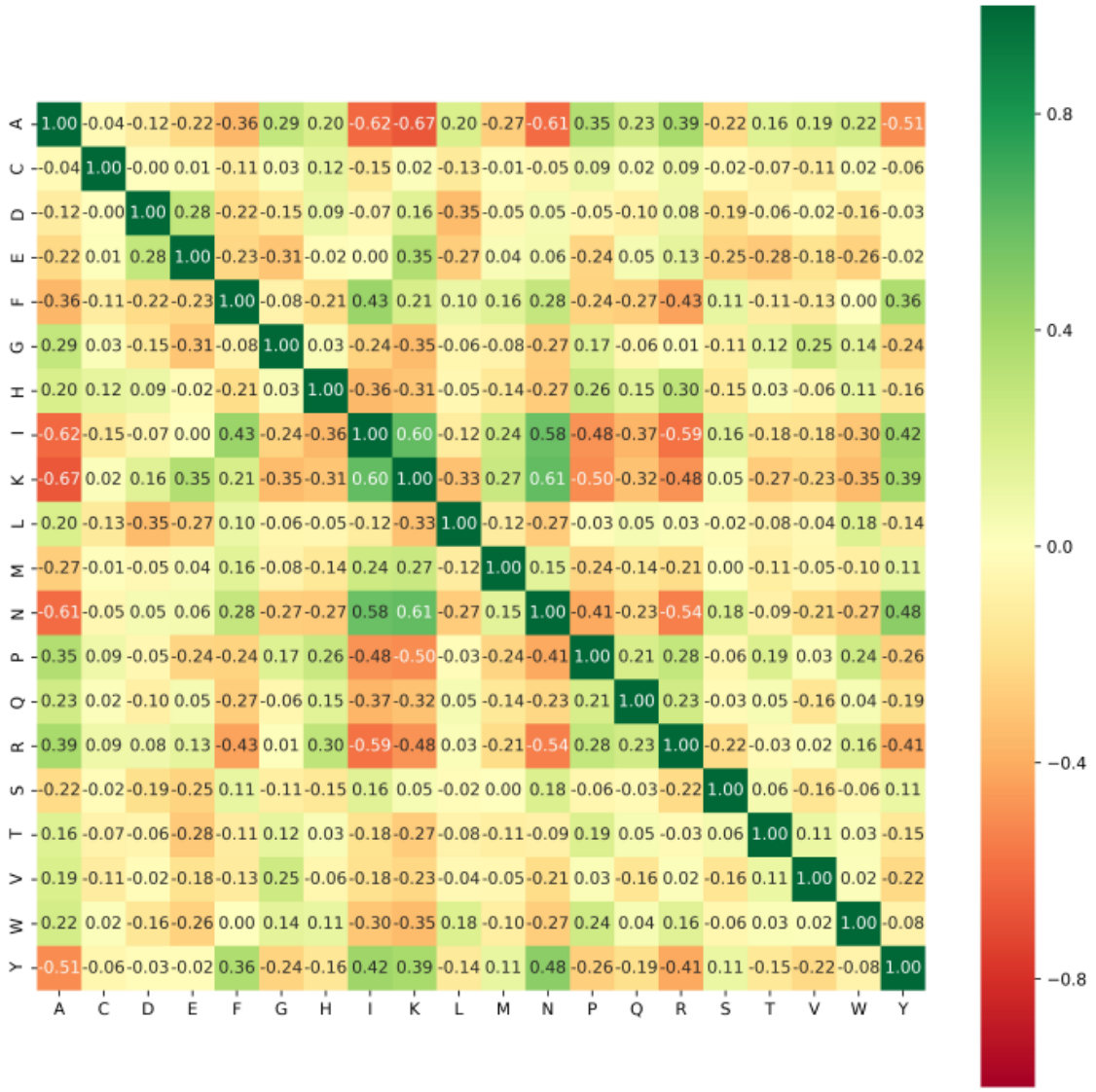


Figure 6.5 Pearson correlation coefficient between amino acid frequencies in thermophilic proteins.

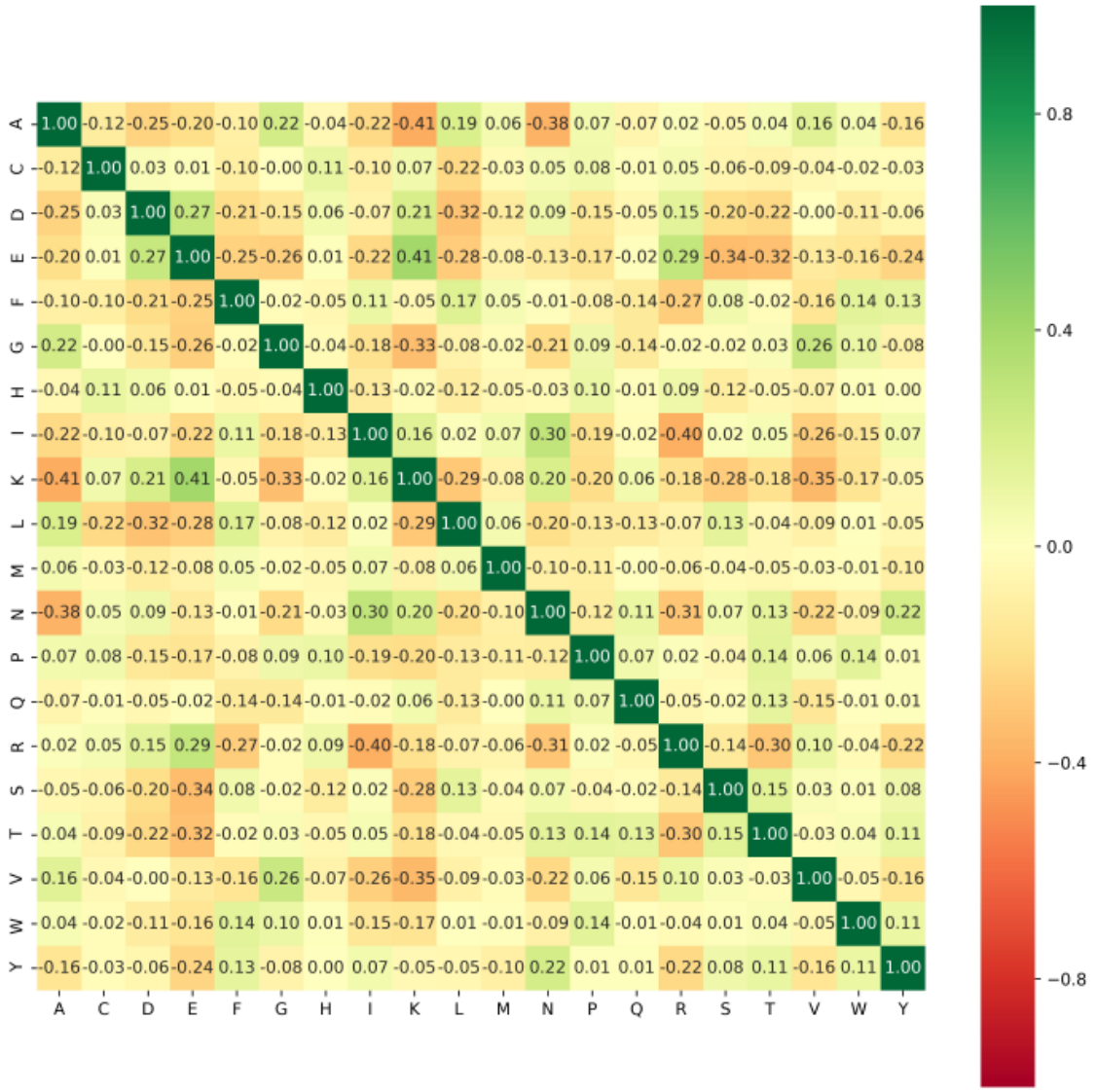


Figure 6.6 Pearson correlation coefficient between amino acid frequencies in hyperthermophilic proteins.

6.4.5 Differences in amino acid correlation

Having determined the inter-correlation of amino acid occurrence in proteins, we tested the hypothesis:

$$H_o: r_{x,y}^{S} = r_{x,y}^{\overline{S}}$$

$$H_a: r_{x,y}^{S} \neq r_{x,y}^{\overline{S}}$$

First, the correlation coefficients were transformed to a normal distribution using Fisher transforms and *p*-values were determined from the calculated Z-scores. We rejected the null hypothesis for $p < 0.01$. From the results, 63 amino acid pairs have significantly different correlation coefficients (Table 6.10). A positive difference in the correlations of amino acids, x and y, between the set, S, and $\overline{S}$ implies that x and y are much more positively constrained in S than they are in $\overline{S}$. For example, Pro and Thr (P,T) are more positively constrained in thermophilic and hyperthermophilic proteins than they are in other proteins, and this constraint is higher in thermophilic proteins than in hyperthermophilic proteins.

Table 6.10 Differences in correlation coefficient of amino acid frequencies for psychrophilic, mesophilic, thermophilic and hyperthermophilic proteins. P - MTH is the difference between the correlation coefficient of amino acids in psychrophilic proteins and the correlation coefficient in other proteins (M, T, and H). P-MTH, M-PTH, T-PMH, and H-PMT are significant (p<0.01) for all 63 amino acid pairs.

S/No.	Amino acid pair	P-MTH	M-PTH	T-PMH	H-PMT
1	R, N	0.349	0.174	-0.198	0.169
2	Q, K	0.156	-0.174	-0.198	0.233
3	A, R	-0.310	-0.144	0.376	-0.285
4	M, A	0.184	0.159	-0.343	0.148
5	A, K	0.235	0.088	-0.253	0.157
6	K, H	0.197	-0.114	-0.209	0.140
7	K, M	-0.111	-0.152	0.323	-0.189
8	R, F	0.185	0.092	-0.168	0.093
9	Q, I	0.100	-0.194	-0.090	0.296
10	I, L	0.199	-0.097	-0.113	0.103
11	H, N	0.092	-0.144	-0.223	0.105
12	I, F	-0.140	0.083	0.227	-0.243
13	W, N	0.096	-0.087	-0.107	0.137
14	P, T	-0.091	-0.089	0.163	0.095
15	N, M	-0.222	-0.132	0.296	-0.089
16	K, R	0.424	0.077	-0.324	0.221
17	W, L	-0.108	0.133	0.082	-0.153
18	Q, N	0.083	-0.333	-0.295	0.193
19	N, K	-0.358	0.067	0.345	-0.284
20	I, R	0.297	0.077	-0.298	0.066
21	M, Y	-0.106	-0.090	0.216	-0.080
22	K, W	0.157	-0.069	-0.140	0.116
23	H, R	-0.066	0.131	0.235	-0.069
24	Q, T	-0.131	-0.145	-0.067	0.116
25	P, R	-0.254	-0.063	0.141	-0.330
26	E, I	-0.089	0.065	0.117	-0.159
27	R, E	-0.058	-0.151	-0.142	0.134
28	W, F	0.097	-0.059	-0.091	0.106
29	W, T	-0.065	-0.062	0.059	0.073
30	P, W	-0.098	0.064	0.055	-0.073
31	M, I	-0.121	-0.084	0.203	-0.057
32	R, V	-0.137	-0.325	-0.091	0.057
33	P, M	0.173	0.105	-0.178	0.056
34	M, R	0.139	0.056	-0.155	0.104

Table 6.10 (continued)

S/No.	Amino acid pair	P-MTH	M-PTH	T-PMH	H-PMT
35	Q, A	-0.119	0.107	0.053	-0.213
36	W, E	0.057	-0.062	-0.078	0.053
37	Y, P	0.154	-0.053	-0.202	0.220
38	T, S	-0.073	-0.054	-0.086	0.063
39	F, H	0.052	-0.064	-0.131	0.076
40	S, K	-0.054	0.052	0.165	-0.330
41	W, R	-0.101	0.051	0.127	-0.163
42	I, V	0.050	0.159	-0.103	-0.133
43	Y, K	-0.189	0.047	0.301	-0.314
44	E, N	0.047	0.224	0.135	-0.159
45	W, I	0.177	-0.045	-0.131	0.079
46	Y, L	0.050	-0.084	-0.046	0.091
47	L, T	-0.079	-0.050	0.046	0.100
48	Y, R	0.241	0.084	-0.319	0.043
49	H, P	-0.088	0.051	0.168	-0.045
50	Q, F	-0.042	-0.201	-0.094	0.109
51	E, S	0.099	0.143	0.040	-0.119
52	Q, V	0.123	0.104	0.140	0.041
53	G, A	-0.071	0.063	0.037	-0.106
54	L, Q	0.040	0.115	0.053	-0.178
55	V, A	0.081	-0.074	0.099	-0.037
56	A, D	0.064	-0.036	0.040	-0.092
57	W, Y	0.137	-0.037	-0.153	0.112
58	K, I	-0.388	-0.032	0.398	-0.256
59	N, Y	-0.205	0.032	0.315	-0.112
60	N, C	-0.035	-0.081	-0.058	0.069
61	D, V	-0.033	-0.041	0.115	0.107
62	T, F	-0.060	-0.033	-0.059	0.064
63	M, L	0.033	0.045	-0.149	0.100

6.5 Discussion

We have tested machine-learning algorithms in predicting the thermostability of 32,000 diverse proteins with less than 40% sequence identity. In our 5-fold cross validation tests, the RBF SVM method demonstrates higher predictive accuracy than other methods in discriminating psychrophilic, mesophilic, thermophilic, and hyperthermophilic proteins. So that researchers are able to access these tools, we make the RBF SVM classifiers available as a Python module (ThermoProt) at https://github.com/jafetgado/ThermoProt.

In all methods we tested, the lowest accuracy was achieved in the discrimination of psychrophilic from mesophilic proteins (PM), with differences in accuracy of as much as 11.4% compared with the discrimination of other protein classes (Table 6.5). This suggests that in amino acid composition and, consequently, structural features arising from the amino acid distribution (such as hydrophobicity), mesophilic proteins are more similar to psychrophilic proteins than thermophilic proteins. Hence, fewer and less drastic modifications may be required to adapt a mesophilic protein to low temperature environments than to higher temperature environments.

In this work, 50 features were employed in our machine-learning models. From the random forest classifiers, we compared the Gini importances of these features (Figure 6.7). From the Gini importance, we observed that dipeptide composition (DPC) features are the least important features. Only four DPC features are among the top 25 features (EQ 0-gap, RR 0-gap, KQ 0-gap, and L**Q 2-gap). This indicates that there is more discriminatory information in the amino acid composition than in dipeptide composition. The heat capacity is the most important feature for the MT classifier and is among the top four features for the MTH and TH classifiers. Interestingly, although heat capacity correlates

weakly with thermostability (Table 7.3) and no significant trend was observed in the distribution of heat capacity for the different thermostability classes (Figure 6.8), the heat capacity is a powerful feature when used in combination with other features.

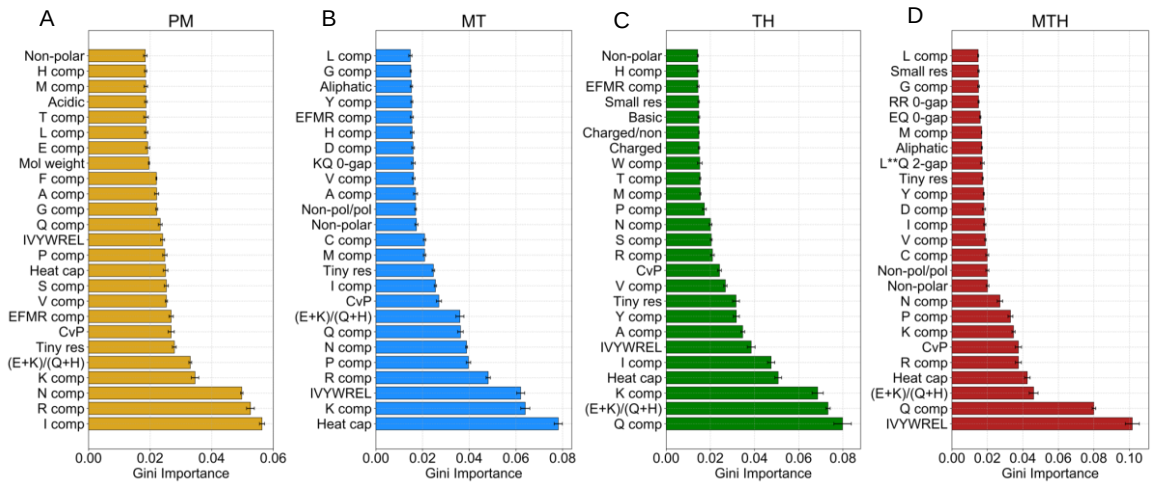


Figure 6.7 Relative (Gini) importance of top 25 features in random forest discrimination of (A) Psychrophilic vs. mesophilic (PM) proteins (B) Mesophilic vs. thermophilic (MT) proteins (C) Thermophilic vs. hyperthermophilic (TH) proteins (D) Mesophilic vs. thermophilic and hyperthermophilic (MTH) proteins. See Table 6.3 for full description of features along the y-axis.

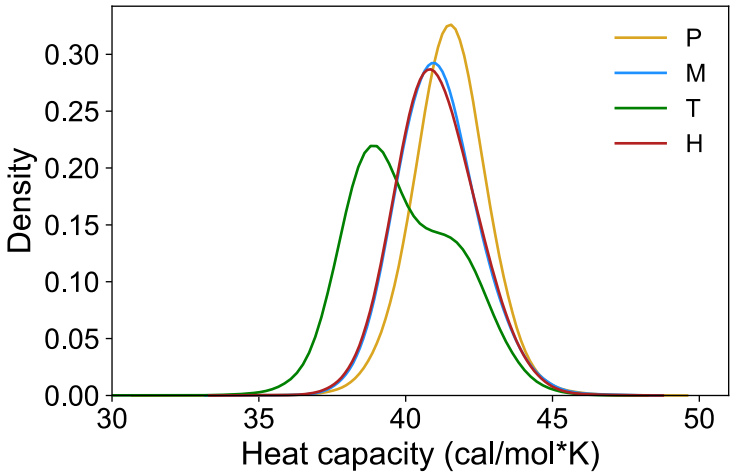


Figure 6.8 Distribution of heat capacities for 32,000 psychrophilic (P), mesophilic (M), thermophilic (T), and hyperthermophilic (H) proteins. The heat capacity of each protein sequence was obtained by a mole-weighted sum of the heat capacities of the constituent amino acids.

The relationships between amino acid composition and protein thermostability have been discussed in previous research.[369, 370, 372-374] However, many of such studies have used relatively small datasets so that the trends observed may not represent the wide variety of proteins in the databases. In this work, we investigated the relationships between amino acid composition and thermostability by measuring the Spearman's rank correlation coefficients between amino acid composition and the thermostability class (i.e. P=1, M=2, T=3, H=4) of 32,000 proteins (Figure 6.9). Spearman's rank correlation was chosen as the preferred correlation measure since it is non-parametric and is more appropriate for ordinal data. Higher positive values of the Spearman's coefficient indicate that the amino acid is preferred by thermostable proteins and likely plays an important role in enhancing thermal stability.

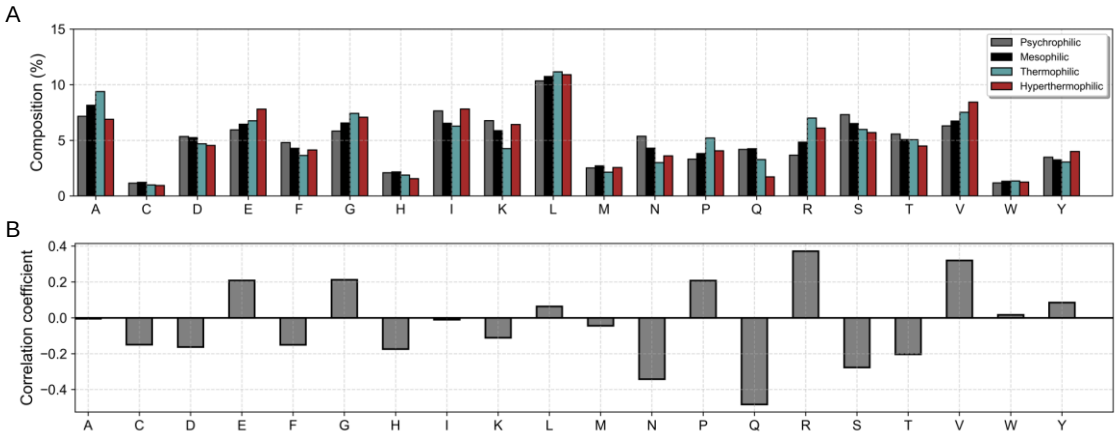


Figure 6.9 Relationship between thermal stability and amino acid composition for 32,000 proteins used in validation tests.(A) Average amino acid composition of psychrophilic, mesophilic, thermophilic, and hyperthermophilic proteins. (B) Spearman's correlation coefficient between amino acid composition and thermostability class (i.e. P=1, M=2, T=3, H=4). Correlation (r_s) is significant ($p < 0.01$) for all $|r_s| > 0.015$.

It has been observed that higher amounts of charged amino acids are employed by thermophilic proteins to form stabilizing salt bridges and hydrogen bonds, and thermophilic proteins have a preference of Arg and Glu over Asp and Lys.[367, 373, 374, 398] Our results indicate that, whereas there is a strong positive correlation between thermostability and the compositions of Arg and Glu, there is a negative correlation with Lys and Asp. Moreover, Asp is known to be unstable at high temperatures and tends to be avoided in thermophilic proteins.[374, 399] We also observed that all uncharged polar amino acids negatively correlate with thermostability, with Gln having the strongest negative correlation. Gln, as well as Asn, Met, and Cys are known to be less frequent in thermophilic proteins since they are thermolabile.[400, 401] Although disulfide bridges improve

thermostability, the negative contribution due to the thermolability of Cys may outweigh the positive contribution due to disulfide bridges, such that Cys correlates inversely with thermostability. Furthermore, the uncharged polar amino acids (Ser, Thr) tend to be replaced with charged amino acids in thermophilic proteins to maximize the formation of hydrogen bonds and salt bridges.

Tyr is the preferred aromatic amino acid in thermophilic proteins; this may be due, in part, to the hydroxyl group which allows for hydrogen bonding,[372] and Tyr tends to form more amino-aromatic contacts, resulting in better packing in the core of the protein.[402] We observed that Tyr correlates positively with thermostability, Trp concentration is similar across all proteins, and Phe and His correlate inversely with thermostability. Thermophilic proteins tend utilize a higher composition of hydrophobic residues to enhance rigidity due to hydrophobic interactions.[375, 376] Thus, we observe that Val, Leu, Gly, and Pro correlate positively with thermostability, although Ala and Ile show no significant trend. Moreover, Gly and Pro composition is lower in hyperthermophiles compared to thermophiles. It has been suggested that psychrophilic proteins have an increased concentration of Gly to provide for improved conformational mobility.[403-405] On the contrary, our results suggest that this is not the case. Rather, psychrophilic proteins have the lowest Gly content. Perhaps, the stabilizing effects of an increase in hydrophobicity outweigh the increase in flexibility derived from a larger Gly content. To circumvent this, Gly in psychrophilic proteins is more likely to be located in strategic regions that maximize the flexibility.[404]

We also examined pairwise correlations of amino acids in psychrophilic, mesophilic, thermophilic, and hyperthermophilic proteins. Strong correlation coefficients between amino acids indicate a mutual constraint in their occurrence imposed by structural

and, consequently, functional properties. Correlations between amino acids differ significantly in α, β, α/β, and α+β proteins and have been successfully applied to predict the structural class of proteins.[406] Here, we tested the hypothesis that the pairwise correlations between amino acids are significantly different in psychrophilic, mesophilic, thermophilic, and hyperthermophilic proteins. Our results indicate that the correlation between 63 amino acid pairs (out of the 190 possible combinations) are significantly different ($p < 0.01$) (Figures 6.6 to 6.9, Table 6.10). For example, there is a significant positive difference of 0.376 ($p < 0.0001$) in the correlation of Arg and Ala in thermophilic proteins compared to psychrophilic, mesophilic and hyperthermophilic proteins. Hence, Arg and Ala are more constrained in thermophilic proteins than they are in the other protein classes, and larger amounts of Arg imply larger amounts of Ala in thermophilic proteins. This is probably to ensure a balance between increased hydrophobicity due to higher amounts of Ala in the protein core, and increased polar surface composition due to higher amounts of Arg.

6.6 Conclusions

We have applied machine learning to predict the thermal stability of proteins. Our method employs a unique combination of features on a larger and more diverse set of proteins than used in previous studies and performs competitively. We have also distributed our machine-learning models as an easy-to-use Python package (ThermoProt) for efficient prediction of protein thermostability. ThermoProt will find useful applications in high-throughput screening for thermostable proteins particularly when structure data and metadata are unavailable.

CHAPTER 7. Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble learning.

This chapter has been reprinted with permission from Gado et al,[407] copyright 2020, American Chemical Society. The author of this dissertation performed all computational calculations and analyses in this chapter.

7.1 Abstract

Accurate prediction of the optimal catalytic temperature (T_{opt}) of enzymes is vital in biotechnology, as enzymes with high T_{opt} values are desired for enhanced reaction rates. Recently, a machine-learning method (TOME) for predicting T_{opt} was developed. TOME was trained on a normally-distributed dataset with a median T_{opt} of 37˚C and less than five percent of T_{opt} values above 85˚C, limiting the method's predictive capabilities for thermostable enzymes. Due to the distribution of the training data, the mean squared error on T_{opt} values greater than 85°C is nearly an order of magnitude higher than the error on values between 30 and 50°C. In this study, we apply ensemble learning and resampling strategies that tackle the data imbalance to significantly decrease the error on high T_{opt} values (>85˚C) by 60% and increase the overall R^2 value from 0.527 to 0.632. The revised method, TOMER, and the resampling strategies applied in this work are freely available to other researchers as Python packages on GitHub.

7.2 Introduction

Enzymes that are stable and active at high temperatures are especially desirable for industrial applications, as they enable biochemical processes to be conducted at higher temperatures yielding faster reaction rates. Hence, researchers have long sought to develop tools for accurate *in silico* prediction of enzyme thermostability. Accordingly, many tools have been developed over the past two decades to predict the enzyme melting temperature (T_m),[408-410] the change in thermodynamic stability ($\Delta\Delta G$) upon point mutations,[411-419] or the optimal growth temperature (OGT) of the source organism.[381, 387, 388, 420-425] Unfortunately, for prediction purposes, higher OGT or thermal stability do not necessarily indicate substantial catalytic activity at high temperatures.[368, 426] Hence, a tool that directly predicts the optimal catalytic temperature (T_{opt}) of enzymes is desirable.

Recently, Li *et al.* developed a machine-learning tool, TOME (Temperature Optima for Microorganisms and Enzymes), for predicting the OGT of microorganisms and the T_{opt} of enzymes.[426] TOME uses a support vector regressor to predict OGT from the dipeptide composition of the proteome, and a random forest regressor to predict T_{opt} from the OGT and the amino acid composition. In predicting OGT, TOME achieved an R^2 value of 0.88 in cross validation tests, which is superior to other published models.[427, 428] However, the R^2 value of T_{opt} prediction was only 0.51, providing impetus for further improvement. More recently,[429] Li *et al.* incorporated feature engineering to improve the accuracy of T_{opt} prediction. They extracted 5,494 and 5,700 sequence features, using the packages, iFeature and UniRep, respectively.[430, 431] However, these features did not provide a significant improvement in performance compared to using only the amino acid composition and OGT, even when deep learning was applied. As a result, the authors concluded that more

informative features, such as features from the three-dimensional structure, may be necessary to markedly improve T_{opt} prediction performance. Yet, a tool that accurately predicts T_{opt} from sequence-data alone remains valuable to the biotechnology community, since it can be readily applied to the vast number of proteins in the databases that lack structural characterizations.

In this work, we sought to improve the accuracy of T_{opt} prediction, not by customary feature engineering, but by mitigating the adverse impact of the non-uniform distribution of the training data used in the machine learning model. It is recognized that an imbalanced data distribution is highly unfavorable in machine learning problems, as it biases the learning algorithms towards the abundant data regions at the expense of the poorly sampled regions, and, thus, leads to higher error on the rare values and overall sub-optimal model performance.[113, 114, 432] In classification problems, data imbalance has been extensively studied, and numerous techniques for dealing with imbalance problems have been proposed.[115, 433] These methods are generally classified into three groups: algorithm-level methods, which specifically modify the learning algorithm to address the bias; data-level methods, which resample the data in a preprocessing step to decrease the unevenness of the data; and hybrid methods, which combine both algorithm- and data-level methods.[114, 434] Data-level methods modify the data distribution primarily by either undersampling the majority class, oversampling the minority class, or a combination of both.[434] Researchers have developed multiple resampling methods for classification problems such as neighborhood cleaning rule (NCL),[435] synthetic minority oversampling technique (SMOTE),[116] selective preprocessing of imbalanced data (SPIDER),[436] and majority undersampling technique (MUTE).[437] The combination of resampling strategies with

ensemble learning (the integration of the outcomes of multiple base models) has proven remarkably successful in dealing with class imbalance.[434, 438, 439]

On the contrary, less attention has been paid to imbalance in regression problems.[114, 115] Few methods have been proposed for working with imbalanced distributions in regression domains including: SMOTE for regression (SMOTER),[122] SMOGN,[120] meta learning for utility maximization (MetaUtil),[440] resampled bagging (REBAGG),[121] and weighted relevance-based combination strategy (WERCS).[118] In many bioinformatic and cheminformatic supervised-learning regression problems, the data often follows a normal distribution, and the rare extreme values may be more important to the user than the abundant values centered about the median of the distribution. For example, in predicting T_{opt} for practical applications, higher T_{opt} values are generally more relevant since thermostable enzymes are desired for enhanced biochemical reaction rates. Still, a majority of studies do not address the issue of data imbalance,[417, 418, 441, 442] resulting in models with reduced predictive accuracy at tails of the normal distribution.[115, 118] Additionally, standard metrics used in assessing regression model performance, such as mean squared error (MSE) and mean absolute deviation (MAD), are heavily biased towards the abundant values centered about the median so that the reported performance fails to capture the poorer performance on rare values at the tails of the distribution.[443] Consequently, a model could demonstrate excellent performance on non-uniform datasets and, yet, have little ability to accurately predict extreme values.

In this study, we apply resampling and ensemble methods to enzyme T_{opt} prediction. Our results show that without resampling (i.e., TOME), the error (MSE) in predicting high temperature values (>65°C) was about 500% higher than the error in

predicting T_{opt} values centered about the median (30-50°C). By applying resampling strategies alone, without the introduction of new features, we were able to reduce the error on high temperature values (>65°C) by more than 50% and, consequently, increase the overall performance (R^2) by 20%. We make available the machine-learning tool for improved T_{opt} prediction, TOMER (Temperature Optima for Enzymes with Resampling), through GitHub. We anticipate TOMER will prove valuable in accurately predicting T_{opt} values of industrially-relevant, thermostable enzymes. To facilitate minimizing the impact of data imbalance in other regression applications, we have also provided the resampling strategies employed here as a Python package, resreg (Resampling for Regression).

7.3 Methods

7.3.1 Dataset and machine learning implementation

The dataset used in training TOMER was obtained from Li *et al.*, consisting of 2,917 enzymes with experimental T_{opt} measurements and OGT data from the BRENDA database.[429, 444] Throughout this work, all machine learning regressors were trained on the same 21 features used in TOME, which include the frequencies of the 20 amino acids and the OGT. The features were normalized by subtracting the mean and dividing by the standard deviation before fitting the regressors. Machine learning was implemented with the scikit-learn package (v0.21.2)[445] in Python (v3.6.6).

7.3.2 Evaluation of performance

In evaluating the performance of the regressors, we did not use the conventional *k*-fold cross validation technique. Since the data are normally distributed, randomly splitting

the data into folds will result in similarly imbalanced folds and, as a result, the performance metrics (R^2, MSE) will overly weight the frequent data and will not sufficiently capture the performance at the distribution tails. Hence, we evaluated performance of the regressors on a testing set that was nearly uniformly distributed. A uniform testing set was formed by splitting the entire dataset into five bins based on the target values (T_{opt}). Then, 70 samples were randomly selected from each bin to constitute the testing set, with the remaining data forming the training set (Table 7.1, Figure 7.1A). We selected only 70 samples from each bin so that at least half of the data in the smallest bin (85-120°C) was used in training. This way, 88% of the entire dataset was used in training (2,567 samples) and 12% in testing (350 samples). The dataset was repeatedly split into training and testing sets 50 times, and each time, resampling strategies were applied to the training set before fitting the regressors. The performance on the testing set was measured as an average over the 50 iterations, i.e., Monte Carlo cross validation (MCCV).[446]

Table 7.1 Formation of a uniform testing set by selecting equal samples from five bins.

Bins	Range (°C)	Samples in bin	Percent of total dataset	Testing size	Training size
0-30	$0 \leq y < 30$	461	15.8%	70	391
30-50	$30 \leq y < 50$	1427	48.9%	70	1357
50-65	$50 \leq y < 65$	519	17.8%	70	449
65-85	$65 \leq y < 85$	361	12.4%	70	291
85-120	$85 \leq y \leq 120$	149	5.1%	70	79
	Total	2,917	100%	350 (12%)	2,567 (88%)

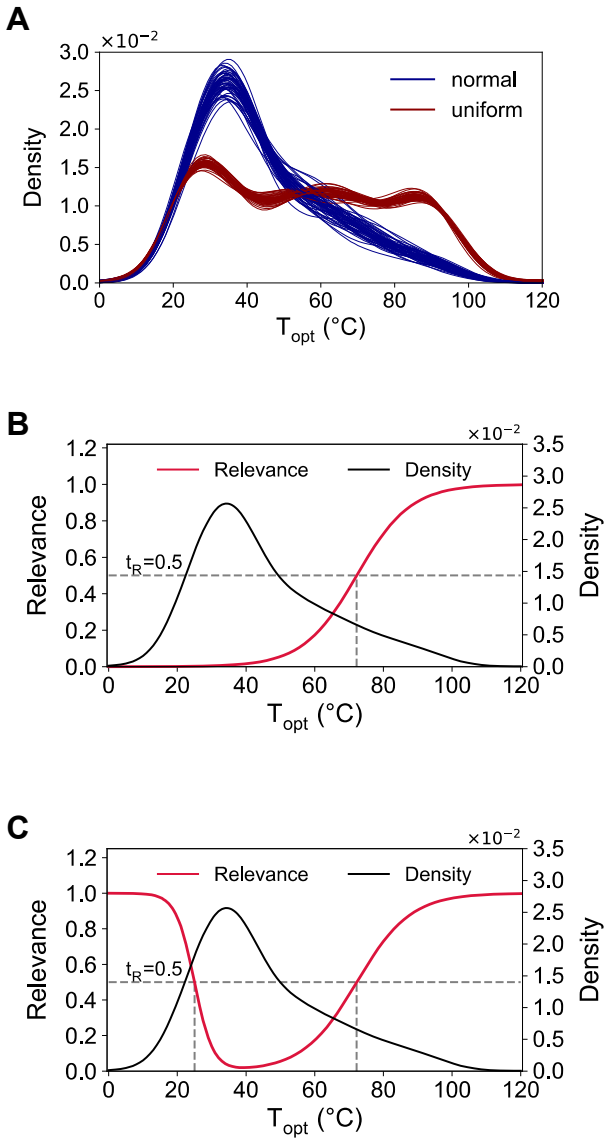


Figure 7.1 Distribution of Topt values in the dataset of 2,917 proteins. The density plots were derived using a Gaussian kernel density estimation (KDE). (A) Distribution of testing set in 50 iterations of Monte Carlo cross validation. A normally-distributed testing set formed by random selection of 350 samples is shown in blue, and the nearly uniform testing set formed by selecting 70 samples from five bins is shown in red. (B) A one-sided sigmoid relevance function that maps T_{opt} values to relevance values between 0 and 1 (left-hand *y*-

axis). By setting the value of *c* in the relevance function (Equation 7.5) to the 90th percentile (72.2), T_{opt} values greater than 72.2°C form the rare domain (shaded region) and all other values form the normal domain. The T_{opt} distribution density is shown on the right-hand -axis. (C) A two-sided relevance function mapping T_{opt} values to relevance values between 0 and 1. By setting the values of c_L and c_H in the relevance function (Equation 7.6) to be the 10th and 90th percentile (25 and 72.2, respectively), T_{opt} values less than 25°C and greater than 72.2°C form the rare domain, and the complement of the rare domain forms the normal domain. The T_{opt} distribution density is shown on the right-hand *y*-axis.

Four metrics were used to assess the predictive performance. The coefficient of determination (R^2) on a uniformly-distributed test set was used to assess the overall performance, and was the primary metric for selecting the best resampling strategy. Both real and predicted T_{opt} values were converted to categorical values (0-30 is 1, 30-50 is 2, 50-65 is 3, etc., see Table 7.1), and the Matthew's correlation coefficient (MCC)[447] was determined as for a multiclass classification problem.[448] The mean squared error (MSE) was calculated for each bin to evaluate the variation in the performance across the range of T_{opt} values and to examine the error on rare high values relative to the error on abundant values . Finally, we measured the F_1 score as a way to assess the predictive performance on high T_{opt} values at the distribution's tail ($\geq$ 65°C). The F_1 score, which is the weighted harmonic mean of precision and recall, is typically a classification performance metric, but has been adapted for regression problems.[122, 443] For regression, recall and precision has been defined as:[443]

$$precision = \frac{\sum_{\phi(\hat{y}_i) \geq t_R} \left(\alpha(y_i, \hat{y}_i) \times \phi(\hat{y}_i) \right)}{\sum_{\phi(\hat{y}_i) \geq t_R} \phi(\hat{y}_i)} \tag{7.1}$$

$$recall = \frac{\sum_{\phi(y_i) \geq t_R} \left(\alpha(y_i, \hat{y}_i) \times \phi(y_i) \right)}{\sum_{\phi(y_i) \geq t_R} \phi(y_i)} \tag{7.2}$$

where y_i and $\hat{y}_i$ are the true and predicted T_{opt} values, respectively; $\phi(\hat{y}_i)$ is the relevance function which maps the target values to a relevance scale from 0 to 1 (discussed below); t_R is the relevance threshold that forms the subdomain of relevant rare values, and $\alpha(y_i, \hat{y}_i)$ is a function that defines the accuracy of a prediction. Hence, the precision and recall are measures of the predictive accuracy on rare values, weighted by the relevance function. The accuracy function was defined as:[443]

$$\alpha(y_i, \hat{y}_i) = I(L(y_i, \hat{y}_i) \leq t_L) \times \left(1 - exp\left(\frac{-k(L(y_i, \hat{y}_i) - t_L)^2}{{t_L}^2} \right) \right) \tag{7.3}$$

where $L(y_i, \hat{y}_i)$ is the loss function and is equal to the absolute error of the prediction; I is the indicator function, which returns 1 if the absolute error is less than a threshold loss, t_L, but zero otherwise; and k is an integer that defines the steepness of the accuracy curve. We set k to be 10^4 and t_L to be 5 so that predictions within error limits of 5°C are regarded as accurate. A right-sided relevance function was used, with $t_R \geq 0.5$ for all $y \geq 65$ (see Equations 7.5 and 7.6), and the F_1 score was calculated from precision and recall as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{7.4}$$

7.3.3 The relevance function

In classification problems, resampling strategies can be readily applied since the target values are clearly divided into discrete classes. Resampling is not as straightforward

in regression problems, however, since the target variable is continuous. The concept of a relevance function was introduced in previous works to simplify resampling in regression problems.[122, 443, 449] The relevance function is a user-defined function that maps the domain of target values to a scale from 0 to 1, where 1 indicates maximum relevance. By specifying a relevance function, $\phi(y)$, and a relevance threshold, t_R, the domain of target values, D, can be split into two sub-domains: a domain of rare values, D_R, which is of greater importance to the user, and the domain of normal values, D_N (Figure 7.1B and C). Consequently, D_R and D_N can be resampled accordingly.

In this work, we use a sigmoid relevance function defined as:[443]

$$\phi(y) = \frac{1}{1 + exp\big(-s(y - c)\big)} \tag{7.5}$$

where y is the target variable, and *s* and *c* are constants that determine the shape and center of the sigmoid, respectively. By defining s as $\pm\frac{\log(10^4-1)}{|c|}$, it follows that $s > 0$ implies that $\phi(y) \geq 0.5$ for all $y \geq c$, and $s < 0$ implies that $\phi(y) \geq 0.5$ for all $y \leq c$.[443] Hence, c can be specified so that extreme target values beyond c have relevance values above a threshold (t_R) of 0.5 and, thus, form the domain of rare values, D_R. Otherwise stated, $D_R = \{y:\ \phi(y) \geq t_R\}$ and $D_N = \{y:\ \phi(y) < t_R\}$. Equation 7.5 is used to determine $\phi(y)$ in the case that the rare domain is formed from extreme values at the left or right of the normal distribution (one-sided). For a two-sided rare domain formed from both left and right extremes, we define the relevance function as:

$$\phi(y) = \frac{1}{1 + exp\big(-|s_L|(y - c_L)\big)} + \frac{1}{1 + exp\big(|s_H|(y - c_H)\big)} \tag{7.6}$$

where the subscripts, L and H, indicate low and high extreme values, respectively (Figure 7.1B and C).

7.3.4 Resampling strategies

Having defined a relevance function to split the dataset into a rare and normal domain, we tested several resampling methods that alter the lopsidedness of the rare domain, relative to the normal domain. The resampling methods were applied to the training set to mitigate the adverse effects of the data imbalance, and then a random forest regressor with default settings was fitted to the resampled training set. We adapted and implemented the following resampling strategies in this work: random oversampling (RO), introduction of Gaussian noise (GN), synthetic minority oversampling technique for regression (SMOTER), weighted relevance-based combination strategy (WERCS), and WERCS with Gaussian noise (WERCS-GN).[118, 122] We give a brief description of these methods below. The pseudocode of these methods is presented at the end of this chapter.

7.3.4.1 Random oversampling (RO)

With the random oversampling strategy,[118, 119] the rare values are oversampled by duplicating randomly selected data points, while the normal values are left unchanged. The amount of oversampling is to be specified by the user and can significantly affect the results. Branco *et al.* suggested two automatic methods of oversampling: balance and extreme.[118] The balance option oversamples the rare domain so that it is equal in size to the normal domain. The extreme option oversamples the rare domain so that the proportion of the size of the rare domain to the size of the normal domain is reversed. For example, if the normal domain is five times larger than the rare domain, the extreme option oversamples the rare domain so that it is five times larger than the normal domain. Here, we introduced a new automatic oversampling method that is intermediate between balance

and extreme, which we dub "average". According to the method selected, the size of the rare domain after oversampling, $|D_R^{new}|$, is determined from the size of the rare and normal domain before resampling ($|D_R|$ and $|D_N|$, respectively) as follows:

$$\text{balance:} \quad |D_R^{new}| = |D_N| \tag{7.7}$$

$$\text{extreme:} \quad |D_R^{new}| = \frac{|D_N|^2}{|D_R|} \tag{7.8}$$

$$\text{average:} \quad |D_R^{new}| = \frac{1}{2}\left(|D_N| + \frac{|D_N|^2}{|D_R|}\right) \tag{7.9}$$

Additionally, the values of c_L and c_H, which determine the points at which the target value is split to normal and rare values, can have significant effects on the performance. Hence, we implemented a grid search to determine the optimal combination of hyperparameters for the resampling strategies. We defined the hyperparameter space as $c_L \in (25, 30, None)$, $c_H \in (72.2, 60)$, and $method \in (balance, average, extreme)$ (Table 7.2). The values for c_L correspond to the 10th and 20th percentile of T_{opt}, and the values of c_H correspond to the 90th and 80th percentile, respectively. A right-sided rare domain is indicated by $c_L = None$ (Figure 7.1B and C).

Table 7 2 Hyperparameters of resampling strategies tested with a grid search.

Strategy	Hyperparameter range
Random oversampling (RO)[118, 119]	c_L = (25, 30, None), c_H = (72.2, 60), method = (balance, average, extreme)
Synthetic minority oversampling technique for regression (SMOTER)[119, 120, 122]	c_L = (25, 30, None), c_H = (72.2, 60), method = (balance, average, extreme), k = (5, 10, 15)
Introduction of Gaussian noise (GN)[118-120]	c_L = (25, 30, None), c_H = (72.2, 60), method = (balance, average, extreme), δ = (0.1, 0.5, 1.0)
Weighted relevance-based combination strategy (WERCS)[118]	c_L = (25, 30, None), c_H = (72.2, 60), over = (0.5, 0.75), under = (0.5, 0.75)
Weighted relevance-based combination strategy with introduction of Gaussian noise (WERCS-GN)[118-120]	c_L = (25, 30, None), c_H = (72.2, 60), over = (0.5, 0.75), under = (0.5, 0.75), δ = (0.1, 0.5, 1.0)
Resampled bagging with random oversampling (BAGG-RO)[119, 121]	c_L = (25, 30, None), c_H = (72.2, 60), method = (balance, variation), s = (300, 600)
Resampled bagging with SMOTER (BAGG-SMT)[119, 121, 122]	c_L = (25, 30, None), c_H = (72.2, 60), method = (balance, variation), k = (5, 10, 15), s = (300, 600)
Resampled bagging with introduction of Gaussian noise (BAGG-GN)[119, 121]	c_L = (25, 30, None), c_H = (72.2, 60), method = (balance, variation), δ = (0.1, 0.5, 1.0), s = (300, 600)
Resampled bagging with WERCS (BAGG-WR)[118, 121]	c_L = (25, 30, None), c_H = (72.2, 60), over = (0.5, 0.75), under = (0.5, 0.75), s = (300, 600)
Resampled bagging with WERCS-GN (BAGG-WRGN)[118, 119, 121]	c_L = (25, 30, None), c_H = (72.2, 60), over = (0.5, 0.75), under = (0.5, 0.75), δ = (0.1, 0.5, 1.0), s = (300, 600)

7.3.4.2 Synthetic minority oversampling technique for regression (SMOTER)

Applying the SMOTER strategy undersamples the normal values and oversamples the rare values by generating synthetic data points through interpolation between each rare value and a random selection of one of its *k*-nearest neighbors.[119, 120, 122] The feature vector and target value of a synthetic instance, X_2 and y_2, respectively, are determined as follows:[122]

$$X_2 = X_1 + r(X_{nn} - X_1) \tag{7.10}$$

$$y_2 = \frac{y_1 . d_{nn} + y_{nn} . d_1}{d_{nn} + d_1} \tag{7.11}$$

where X_1 is the feature vector of an instance in D_R, X_{nn} is one of *k*-nearest neighbors of X_1, $r \in [0, 1]$ is a random number, y_1 and y_{nn} are the target values of X_1 and X_{nn}, respectively, and d_1 and d_{nn} are the Euclidean distances between X_2 and X_1, and between X_2 and X_{nn}, respectively. The amount of undersampling and oversampling was automatically determined according to the following options:

balance:
$$|D_N^{new}| = |D_R^{new}| = \frac{|D_N| + |D_R|}{2} \tag{7.12}$$

extreme:
$$|D_N^{new}| = |D_R| \tag{7.13}$$
$$|D_R^{new}| = |D_N| \tag{7.14}$$

average:
$$|D_N^{new}| = \frac{1}{2}\left(\frac{|D_N| + |D_R|}{2} + |D_R|\right) \tag{7.15}$$
$$|D_R^{new}| = \frac{1}{2}\left(\frac{|D_N| + |D_R|}{2} + |D_N|\right) \tag{7.16}$$

Optimal hyperparameters were similarly determined by a grid search (Table 7.2).

7.3.4.3 Introduction of Gaussian noise (GN)

The GN strategy is identical to SMOTER in every way except that synthetic points are generated by addition of Gaussian noise rather than interpolation.[118-120] Noise based in $N\big(0, \delta \times std(a)\big)$ is separately added to each feature and to the target value of a rare instance, where $std(a)$ is the standard deviation of the attribute (i.e., feature or target value), and δ is a user-defined parameter that determines the amplitude of the noise.

7.3.4.4 Weighted relevance-based combination strategy (WERCS)

Rather than using a relevance threshold to split the data into rare and normal domains as with the previous strategies, the WERCS strategy uses the relevance values as weights to select data points for undersampling and oversampling.[118] The data are oversampled and then undersampled by selecting instances to be duplicated and instances to be removed, respectively. Selection for oversampling and undersampling is performed using probabilities determined from the relevance function. For each target value in the dataset, y_i, we defined the probability that the value is selected for oversampling or undersampling (p_i^{over} and p_i^{under}, respectively) by Equations 7.17 and 7.18.

$$p_i^{over} = \frac{\phi(y_i)}{\sum_{i=1}^{N} \phi(y_i)} \tag{7.17}$$

$$p_i^{under} = \frac{1 - \phi(y_i)}{\sum_{i=1}^{N} \left(1 - \phi(y_i)\right)} \tag{7.18}$$

Hence, rare values with higher relevance are more likely to be selected for oversampling and less likely to be selected for undersampling. The amount of oversampling and undersampling are hyperparameters to be specified by the user in percent (*over* and *under*, respectively).

7.3.4.5 WERCS with Gaussian noise (WERCS-GN)

We modified the WERCS strategy by adding Gaussian noise to the values selected for oversampling by the WERCS strategy. Hence, with WERCS-GN, oversampling is done with synthetic data, instead of by duplicating data points.

7.3.4.6 Combination of resampling strategies with ensemble learning

Ensemble learning involves training different learners and combining their output to generate a final prediction that is more accurate than the individual learners. Branco *et al.* developed the resampled bagging algorithm (REBAGG) for implementing resampling and bagging in imbalanced regression problems.[121, 130] In this work, we applied an adaptation of the REBAGG algorithm to the prediction of T_{opt} values, by implementing the resampling methods described previously in the REBAGG algorithm (pseudocode is at the end of this chapter).

First, the dataset is split into rare and normal domains, D_R and D_N, using the relevance function, as described previously. Then m models are trained on separately resampled bootstrap samples of s items from the training dataset. Two modes of the REBAGG method are applied: balance or variation mode. In balance mode, an equal number of samples, $s/2$, is randomly drawn from D_R and D_N. In the variation mode, however, $p \times s$ samples are drawn from D_R, and $(1 - p) \times s$ samples are drawn from D_N, where p is a randomly selected number from the set, (1/3, 2/5, 1/2, 3/5, 2/3). Hence, in the variation mode, the m models are trained on data that may contain either fewer, equal, or more rare samples than normal samples. If the number of samples to be drawn from D_R is greater than $|D_R|$, then the extra samples are derived by oversampling the rare domain using RO, SMOTER, or GN, resampling methods as described previously. We refer to the REBAGG method in combination with these resampling methods as BAGG-RO, BAGG-SMT, and BAGG-GN, respectively. A similar combination of REBAGG with WERCS and WERCS-GN (referred to as BAGG-WERCS and BAGG-WRGN) was also implemented.

With BAGG-WERCS and BAGG-WRGN, the data are resampled without splitting into rare and normal domains, as in the WERCS and WERCS-GN methods. Then, *s* samples are drawn from the resampled data for training a model in the ensemble. With these resampled bagging strategies, the resampling step is independently repeated for all m models with replacement. Finally, each model is applied to the testing set, and the final prediction is determined by averaging the predictions of all m models. We used a decision tree regressor with default settings as the base regressor and set m to be 100. Other hyperparameters were optimized based on the values shown in Table 7.2.

7.4 Results and Discussion

7.4.1 Resampling strategies significantly improve predictive performance

In this work, we applied machine learning to predict the T_{opt} of 2,917 enzymes.[426] The target values follow a normal distribution that creates a problem of data imbalance. Although the T_{opt} values range from 0 to 120°C, about half of the values fall within 30 to 50°C, and high temperature data are scarce (Table 7.1). To deal with this data imbalance, we implemented ten strategies that abate the imbalance by resampling the training data. For each strategy, we tested several hyperparameters with a grid search (Table 7.2) and selected the hyperparameter combination that yielded the highest average R^2 value on a uniformly-distributed testing set (Table 7.3). Without resampling the training data (i.e., TOME), the average R^2 value over 50 MCCV iterations was 0.527. However, the best performance of the resampling strategies ranged from 0.567 (RO) to 0.632 (BAGG-RO). Similarly, all resampling strategies yielded significantly higher F_1 scores (>0.178) and MCC values (>0.235) compared to TOME, which had an F1 score of 0.137 and an MCC

score of 0.212 (Figure 7.2). These results demonstrate that the resampling strategies improve the predictive performance on high T_{opt} values (> 65°C), as illustrated by the higher F1 scores, and lead to superior overall performance, as illustrated by the higher R^2 and MCC values. It is important to note that some hyperparameter combinations of the resampling strategies led to a reduction in the predictive performance compared to the model that was trained on non-resampled data (TOME) (Figure 7.3). Hence, it is imperative that one test a sufficiently wide range of hyperparameters to determine the optimal hyperparameter combination.

Table 7.3 Best hyperparameter combination for each resampling strategy yielding the highest R2 values as determined by a grid search.

Strategy	Hyperparameter
RO	c_L =None, c_H=60.0, method=balance
SMOTER	c_L=None, c_H=60.0, method=average, k=10
GN	c_L=None, c_H=72.2, method=balance, δ=0.5
WERCS	c_L=None, c_H=72.2, over=0.5, under=0.5
WERCS-GN	c_L=None, c_H=72.2, over=0.5, under=0.5, δ=0.1
BAGG-RO	c_L=None, c_H=72.2, method=variation, s=600
BAGG-SMT	c_L=None, c_H=72.2, method=variation, k=5, s=600
BAGG-GN	c_L=None, c_H=72.2, method=variation, δ=1.0, s=600
BAGG-WERCS	c_L=25.0, c_H=72.2, over=0.5, under=0.75, s=600
BAGG-WRGN	c_L=25.0, c_H=72.2, over=0.75, under=0.75, δ=0.1, s=600

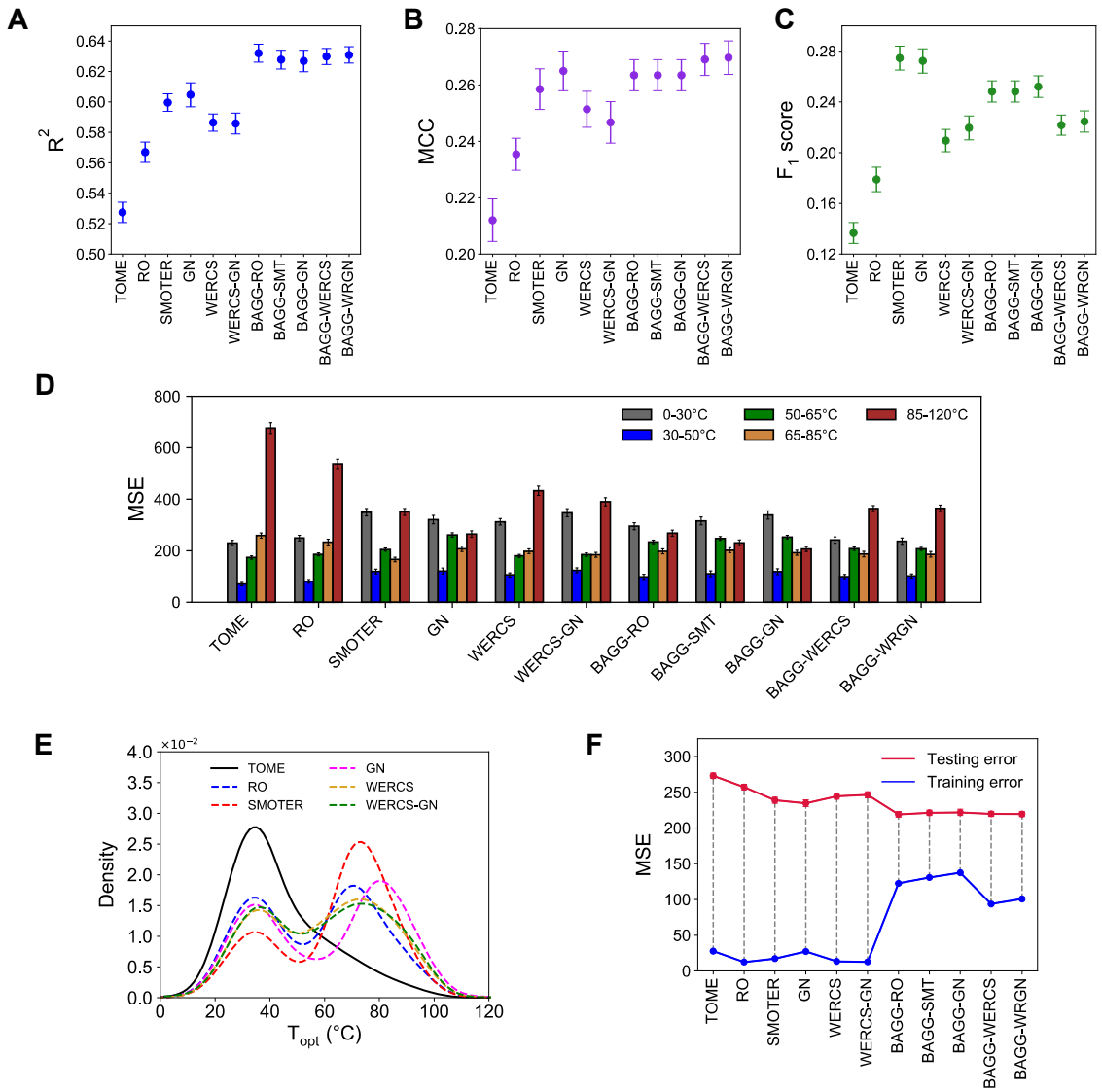


Figure 7.2 Performance of the resampling strategies. The resampling strategies were applied to the training dataset, regressors were fitted on the resampled data, and the performance was evaluated on a uniformly distributed test set with 50 iterations of Monte Carlo cross validation. Error bars indicate 95% confidence interval of the mean over 50 iterations. (A) Highest R^2 value of the resampling strategies determined from a grid search of hyperparameter combinations. Combining bagging with the resampling strategies via the REBAGG algorithm outperforms the resampling strategies alone. See Figure 7.3 for

the performance of all hyperparameter combinations. (B) MCC and (C) F_1 scores of the best hyperparameter combinations of the resampling strategies, i.e., combinations that yielded the highest R^2 value. (D) Mean squared error on different ranges of the target values. Without resampling (TOME), the error is highest in the 85-120°C range, but all the resampling strategies significantly reduce this error. The lowest overall error is achieved by the BAGG-RO strategy. (E) Distribution (KDE) of the dataset after applying the resampling methods with optimal hyperparameters. (F) Mean squared error when regressors trained on resampled data are applied to the training set and the testing set. The integration of resampling strategies with bagging decreases the variance as shown by an increase in training error and decrease in testing error.

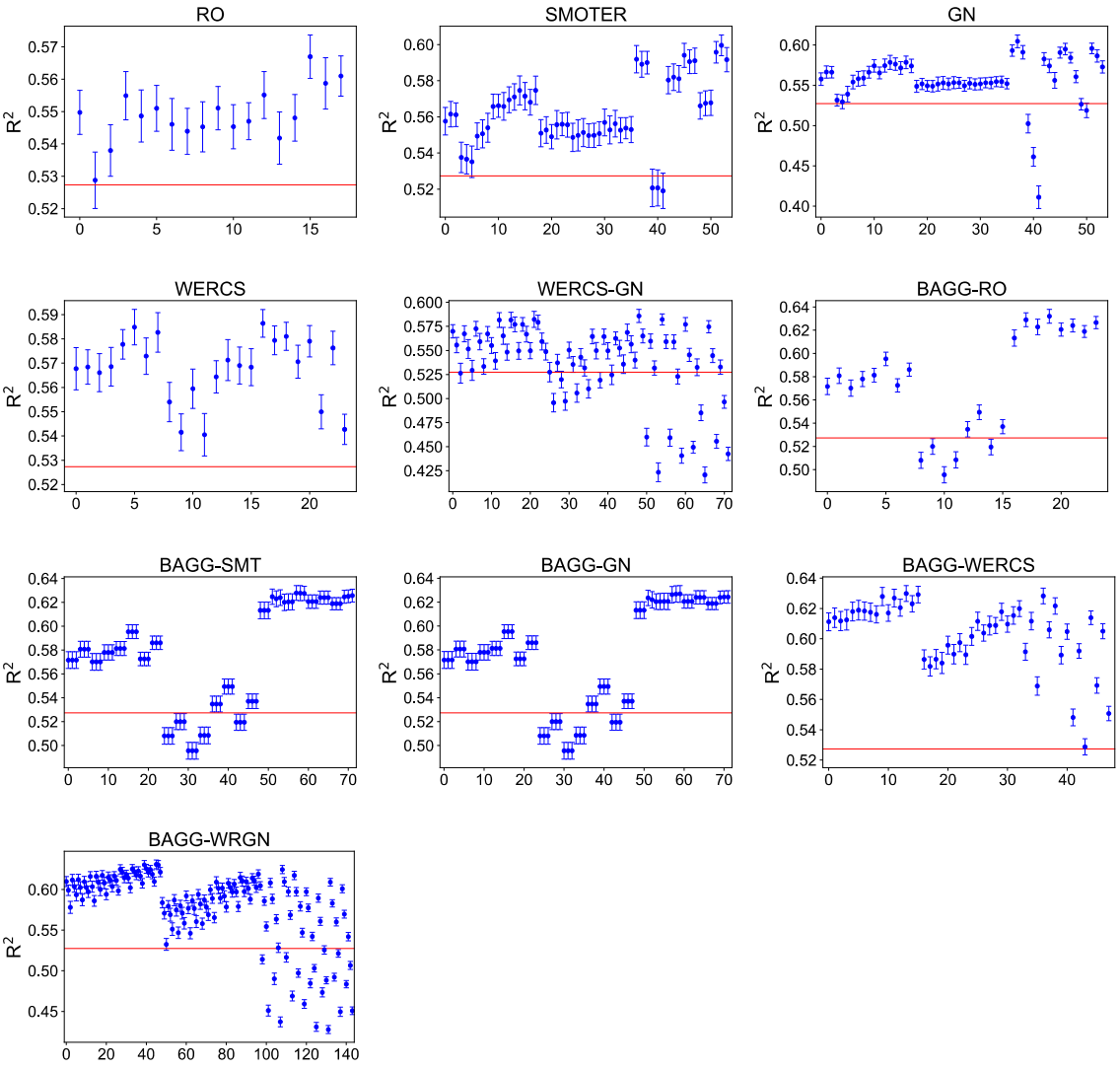


Figure 7.3 Performance (R2) of resampling strategies for all hyperparameter combinations. The *x*-axes indicate different hyperparameter combinations in a grid search. Error bars indicate 95% confidence interval of the mean determined from averaging results over 50 Monte Carlo cross validation repetitions. The red line represents the baseline performance obtained when a random forest regressor is applied to the dataset without resampling (TOME) and the shaded region around the red line indicate the 95% confidence interval. From the figure, it is apparent that some hyperparameter combinations lead to

inferior performance relative to TOME. The BAGG-RO strategy with hyperparameters: C_L=None, C_H=72.2, s=600, and *method*=variation, yielded the highest R^2 value of 0.632.

From the results shown in Figure 7.2A-C, we observed that resampling by simple duplication of rare values, as is done in the random oversampling strategy (RO), led to lower R^2, F_1, and MCC values than the strategies that oversample rare values by using the relevance as weights (WERCS, WERCS-GN), or by generating synthetic data through interpolation (SMOTER) or addition of noise (GN, WERCS-GN). However, this trend was not observed when the resampling methods were combined with bagging (BAGG-RO, BAGG-SMT, BAGG-GN, BAGG-WERCS, BAGG-WRGN). We anticipate that duplication performs worse than generating synthetic values because duplication causes the learning algorithms to overfit to the replicated values. Introducing synthetic values, on the other hand, would cause the algorithms to be more general in the rare data region.[115, 116, 450] Our results indicate that generating synthetic values does not outperform duplication techniques when combined with bagging in the REBAGG strategy, likely because aggregating multiple learners overcomes the overfitting that arises due to replicated values.

Analysis of the MSE as a function of the true T_{opt} values indicates that there is significant variation in the MSE across the range of target values (Figure 7.2D). Without resampling (TOME), the error inversely correlates with the frequency of the data, with lower error in regions of abundant data (30-50°C) and higher error in regions of rare data (0-30°C, 65-120°C). Moreover, error in the 65-85°C and 85-120°C ranges was 3.7 and 9.7 times higher, respectively, than the error in the 30-50°C range. Hence, without resampling, the regressor (TOME) overfits to abundant values and demonstrates inferior performance

on high temperature values. In applications that rely on TOME for identifying high T_{opt} enzymes, the large error on high temperature values may lead to misleading results. By applying resampling strategies to the training set, we altered the distribution of the training dataset to prevent the learning algorithm from overfitting to abundant values and to improve performance on rare high temperature values (Figure 7.2E). As Figure 7.2D shows, all the resampling strategies led to a reduction of the error in the high temperature ranges (65-120°C) and an increase of the error in the abundant data range (30-50°C), which indicates a decrease in the overfitting of abundant values. Moreover, the error in the abundant data range is the lowest error for TOME as well as for all the resampling strategies. This suggests that there is an upper limit to the performance gain from resampling rare data, and more experimental data which sample unexplored regions of the rare data space may be necessary for further improvement in performance.

Furthermore, the combination of resampling methods with bagging, such that each base regressor was trained on independently resampled datasets, yielded significantly higher overall performance scores (R^2 and MCC) than resampling methods alone (Figure 7.2A and B). Other researchers have similarly observed that ensemble learning methods, such as bagging and boosting, considerably enhance the effect of resampling techniques.[434, 438, 451-453] In this work, the resampling methods without bagging (i.e., RO, SMOTER, GN, WERCS, and WERCS-GN) simply increased the proportion of rare values (Figure 7.2E), which decreased the overfitting of the regressors to abundant data, and, consequently, led to a reduction of both training error and testing error (Figure 7.2F). However, the difference between the testing and training error was substantial, indicating that the regressors were overfitting to the resampled training data (high variance). On the other hand, when the

resampling methods were repeatedly applied with multiple decision trees in an ensemble (i.e., the REBAGG strategies) such that each base tree was trained on differently sampled datasets, a much lower testing error and a higher training error was observed. This outcome indicates that the integration of bagging with the resampling methods (i.e., BAGG-RO, BAGG-SMT, BAGG-WERCS, and BAGG-WRGN) reduces the variance of individual regressors and prevents overfitting to the resampled training data, leading to improved generalization.[454] Moreover, all REBAGG strategies yielded similar overall performance (R^2 and MCC), which suggests that the specific resampling method applied in the REBAGG strategy had little effect on the overall performance. The BAGG-RO strategy led to the highest R^2 value of 0.632 and the lowest MSE of 218.6.

7.4.2 Effect of base learners on ensemble performance

We examined the influence of different base learners in the BAGG-RO ensemble to assess whether further performance enhancement could be attained. Using the optimal resampling hyperparameters determined with decision trees (Table 7.3), we applied four additional base regressors in the BAGG-RO ensemble: support vector regressor (SVR), *k*-neighbor regressor (KNR), elastic net (ENET) regressor, and Bayesian ridge regressor (BAYR). For each of these regressors, we used a grid search to determine optimal hyperparameters that yielded the best R^2 value (Table 7.4), and the performance was measured as an average over 50 MCCV iterations. The results indicate that, although each alternative regressor outperformed TOME, the decision tree base regressor yielded the highest R^2 value and lowest overall MSE. Interestingly, the decision tree regressor showed the lowest F_1 score (Figure 7.4). These results suggest that, while other regressors possibly

perform better on high temperature values, tree-based regressors exhibit the best overall performance in predicting T_{opt} values from amino acid composition and OGT.[426, 429]

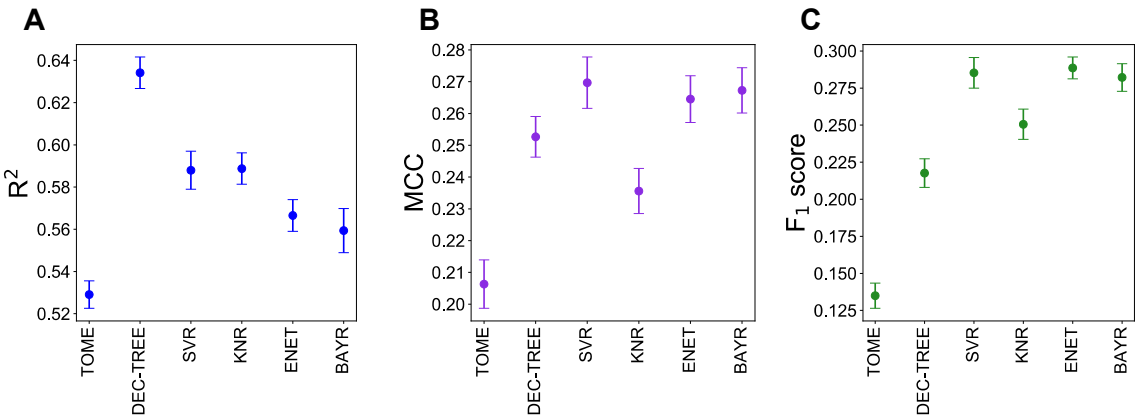


Figure 7.4 Performance of BAGG-RO ensemble with different base learners. The optimal hyperparameters for the base learners were determined by a grid search. Error bars indicate 95% confidence interval of the mean over 50 Monte Carlo cross validation repetitions. (A) Highest R^2 value achieved for different base learners in the BAGG-RO ensemble. (B) Matthew's correlation coefficient and (C) F_1 scores of BAGG-RO strategy with different base learners using the optimal hyperparameters, i.e., hyperparameters that yielded the highest R^2 value.

Table 7.4 Hyperparameters for base learners in BAGG-RO ensemble.

Base learner	Hyperparameter range	Optimal hyperparameters
Support vector regressor (SVR)	C = [10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}], gamma=[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]	C=10^{2}, gamma=10^{-2}
k-neighbor regressor (KNR)	k=[3, 5, 7, 10, 15, 20, 30]	k=3
Elastic net regressor (ENET)	alpha=[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]	alpha=10^{-2}
Bayesian ridge regressor (BAYR)	None (default)	None (default)

7.4.3 Final model, data and code availability

We identified the BAGG-RO strategy with decision tree base learners as the optimal resampling strategy for predicting enzyme optimum temperatures across the entire range of experimental T_{opt} values because it led to the highest R^2 value and lowest overall MSE. A final model was prepared by applying the BAGG-RO resampling strategy with optimal hyperparameters (Table 7.3) to the entire dataset of 2,917 proteins. The final model is available to researchers as a Python package, TOMER (Temperature Optima for Enzymes with Resampling), on the Python package index, http://pypi.org/project/tomer/ with the source code publicly available at http://github.com/jafetgado/tomer/. Compared to TOME, TOMER provides a 20% improvement in the overall predictive performance (R^2), and a 25% and 60% decrease in MSE on T_{opt} values in the 65-85°C and 85-120°C ranges, respectively. All data and code used and produced in this study are available at https://github.com/jafetgado/tomerdesign/. We have also prepared a Python package, resreg (resampling for regression), for applying the resampling strategies discussed in this work to other regression problems. It is available on the Python repository, http://pypi.org/project/resreg, with the source code at http://github.com/jafetgado/resreg.

7.4.4 Pseudocode for resampling strategies as applied in this work

Algorithm 1: Random oversampling algorithm (RO)[118, 119]

Input:

D (X, y) – dataset with features, *X*, and continuous target value, *y*
R – relevance values for corresponding target values
t_R – relevance threshold
over – oversampling percent

Output: *newD* – resampled dataset

rareD ← instances in *D* with relevance $\geq t_R$
normD ← instances in *D* with relevance $< t_R$
addSize ← *over* × |*rareD*| //or determine *addSize* by balance, extreme, or average methods
newRareD ← rareD ∪ randomly selected *addSize* instances from *rareD*
newD ← *normD* ∪ *newRareD*

return *newD*

Algorithm 2: SMOTER algorithm[119, 120, 122]

Input:

D (X, y) – dataset with features, X, and continuous target value, y
R – relevance values for corresponding target values
t_R – relevance threshold
over – oversampling percent
under – undersampling percent
k – number of nearest neighbors

Output: *newD* – resampled dataset

normD ← instances in D with relevance $< t_R$
rareD ← instances in D with relevance $\geq t_R$
lessSize ← *under* × |*normD*|
addSize ← *over* × |*rareD*| //or determine *addSize* and lessSize by balance, extreme, or average methods
rareDL ← instances in *rareD* with $y <$ median(y)
rareDH ← instances in *rareD* with $y \geq$ median(y)
$addSizeL \leftarrow addSize \times \frac{|rareDL|}{|rareD|}$
$addSizeH \leftarrow addSize \times \frac{|rareDH|}{|rareD|}$
nns ← get k-nearest neighbors of all instances in *rareD*
newRareDL ← generate *addSizeL* instances by interpolating between points in *rareDL* and a random selection of one of the k-nearest neighbors
newRareDH ← generate *addSizeH* instances by interpolating between points in *rareDH* and a random selection of one of k-nearest neighbors
newRareD ← *newRareDL* ∪ *newRareDH* ∪ *rareD*
newNormD ← *normD* \ random selection of *lessSize* instances from *normD* //undersampling
newD ← *newNormD* ∪ *newRareD*

return *newD*

Algorithm 3: Introduction of Gaussian noise algorithm (GN)[118-120]

Input:

D (X, y) – dataset with features, X, and continuous target value, y
R – relevance values for corresponding target values
t_R – relevance threshold
over – oversampling percent
under – undersampling percent
δ – magnitude of Gaussian noise

Output: *newD* – resampled dataset

normD ← instances in D with relevance $< t_R$
rareD ← instances in D with relevance $\geq t_R$
lessSize ← *under* × |*normD*|
addSize ← *over* × |*rareD*| //or determine addSize and lessSize by balance, extreme, or average methods
newRareD ← random selection of *addsize* instances from *rareD*

foreach *case* in *newRareD* **do** //add Gaussian noise to each case
    foreach a in $X \cup y$ **do**
        case[a] = *case*[a] + N(0, δ×std(a))
    end
end
newNormD ← *normD* \ random selection of *lessSize* instances from *normD* //undersampling
newD ← *newNormD* ∪ *newRareD*

return *newD*

Algorithm 4: Weighted relevance combination strategy (WERCS and WERCS-GN) algorithm[118]

Input:

D (X, y) – dataset with features, *X*, and continuous target value, *y*
R – relevance values for corresponding target values
over – oversampling percent
under – undersampling percent
δ – magnitude of Gaussian noise

Output: *newD* – resampled dataset

underSize ← *under* × |*D*|
overSize ← *over* × |*D*|
pOver ← $\left\{\frac{r_i}{\sum r_i} \mid r_i \in relevance\right\}$
pUnder ← $\left\{\frac{1-r_i}{\sum 1-r_i} \mid r_i \in relevance\right\}$
underD ← sample *underSize* instances from *D* with *pUnder* weights
newD ←*D* \ *underD* //undersample
overD ← sample *overSize* instances from *D* with *pOver* weights

if method is WERCS-GN **do** //add Gaussian noise
    foreach *case* in *overD* **do**
        foreach *a* in *X* ∪ *y* **do**
            case[*a*] = *case*[*a*] + N(0, δ×std(*a*))
        end
    end
end
newD ←*newD* ∪ *overD*

return *newD*

7.5 Conclusions

In this study, we applied resampling strategies to improve the performance of predicting enzyme optimum temperatures with machine learning. The resampling strategies were implemented to modify the imbalanced distribution of the training set and improve performance on regions with sparse data. Compared with TOME, which at the time of this study is the only available machine-learning tool for predicting enzyme optimum temperatures, our method (TOMER) yields a significant improvement in

predictive accuracy, particularly in the thermophilic regimes. We expect that TOMER will find useful application in high-throughput prospecting of enzymes that are both stable and active at high temperatures. TOMER requires the user to provide the amino acid sequence of the enzyme and the OGT of the source organism. If the OGT is unknown, it may be predicted using TOME.[426] For future considerations, the incorporation of higher-level features or the addition of more experimental data may prove useful strategies for further improving the performance of TOMER. Ultimately, this study highlights the critical need to consider data imbalance in regression problems, especially when the rare, extreme data range is of greater scientific interest than the abundant data region. We anticipate that our Python tool for readily implementing resampling strategies in regression problems (resreg) will be a valuable resource for other researchers in dealing with the challenges of data imbalance.

CHAPTER 8. Conclusions and Future Directions

8.1 Overview

This dissertation investigated the application of data mining and statistical sequence analysis tools to gain functional insight into family 7 glycoside hydrolases, GcoA, MHETase, and to develop predictive models of enzyme thermostability.

Experimental studies by collaborators were done to test the hypotheses generated by the data-driven studies of GcoA and MHETase. In GcoA, conservation analysis revealed the notably variable F169 position within 6Å of the bound guaiacol substrate in the active site. This highlighted F169 as a viable position for protein engineering to expand the substrate specificity of GcoA. Moreover, MD simulations by performed by collaborators observed a stearic clash between the F169 residue and the methoxy group of the non-native substrate, syringol. Mutagenesis and biochemical characterization show that the F169A, unlike the wild type, is markedly active on syringol. This work (Chapter 4) demonstrates the value of data-driven insights into amino-acid variation in an enzyme family for understanding and manipulating function.

In Chapter 5, conservation analysis was combined with phylogenetic analysis to gain insight to key positions in MHETase active site that are crucial for MHET-hydrolase activity. From bioinformatic and phylogenetic analysis, two close homologs were identified, which along with MHETase are strikingly dissimilar from other tannase-family sequences. The results from this study highlight F415 and S131 as positions that specifically evolved in MHETase to accommodate MHET. Experimental studies from collaborators confirm that the F415S and S131G mutants, as well as the two close

homologs, which both lack Phe and Ser at positions 415 and 131, respectively, demonstrate a substantial reduction in MHET-hydrolase activity, relative to MHETase wild type.

Machine learning identified four active-site loops (A4, B2, B3, B4) that strongly relate to functional subtype across the family 7 glycoside hydrolase family. Although previous studies by other researchers have highlighted some of these loops and the striking differences in the structural architecture between cellobiohydrolases and endoglucanases due to their differing lengths, this work is the first attempt investigate the relationships between loop length and functional subtype with a statistical approach and on a large set of sequences. Data mining on the sequence dataset also identified positions that generate classification rules which discriminate GH7 cellobiohydrolases and endoglucanases with high accuracy. Many of these positions have been experimentally studied by other researchers in *Trichoderma reesei* Cel7A and confirmed to play key roles in processive action and substrate binding. Similarly, positions in the catalytic domain that correlate with the presence of a carbohydrate binding module were identified. Altogether, this study provides a practical demonstration of the use of machine learning and data mining techniques to map amino-acid sequence to functional variation in an enzyme family, and, consequently, identify residues that play crucial roles in function.

In addition to enzyme catalytic activity, this work (Chapter 6 and 7) investigated the use of machine learning algorithms to predict protein thermostability. It is particularly interesting that machine learning algorithms can discriminate thermophilic proteins from non-thermophilic proteins with reasonable accuracy by only considering global features derived from the amino-acid sequence (Chapter 6). In Chapter 7, a regression model was built to predict the optimal catalytic temperature, which is much more definite than

classifying enzymes according to their organism growth temperature. By specifically considering the non-uniform distribution of the training data, and implementing resampling strategies to abate the uneven distribution, a strikingly higher predictive performance was achieved, mainly due to improvement at the tail ends of the normally distributed target values. This work demonstrates critical need to deal with data imbalance in regression problems, the ability of machine learning to predict enzyme biochemical kinetic, or thermodynamic properties, and the potential in high-throughput application of machine learning for enzyme discover.

8.2 Future directions

This work focused on mapping protein sequence and structural features to functional properties with supervised learning. Labeled biochemical data is expensive and difficult to generate. Future work will consider unsupervised learning, where the functional labels are scarce or absent. Clustering algorithms will be used to learn the inherent structure and deconstruct the hidden patterns in the sequence data. Representative sequences from each of clusters will be selected, and homology modeling and molecular dynamics simulations will be implemented to gain insight into the structural, dynamic, and possibly, functional meanings of each cluster.

The strategies for dealing with data imbalance in this dissertation were all data-level methods, which resample the data to change the non-uniform distribution. In future studies, algorithm-level methods, which modify the machine learning algorithm to address the non-uniformity in the data, will be considered. Monte Carlo simulations on various

benchmark datasets will be implemented to examine which of these methods yield superior performance and whether a method is more favorable for particular data structures.

Finally, the use of deep learning algorithms to directly predict protein thermostability from the primary structure without feature engineering will be considered. Transfer learning will be used to overcome the challenge of a small training dataset. I hypothesize that pretraining a deep neural network on organism growth temperature data, which is abundantly available, and transferring the trained model to other protein thermostability problems with scarce data will yield superior performance compared to the traditional machine learning approaches applied in this work.

Appendices

A1 Supporting Information for Machine Learning Reveals Sequence-Function Relationships in Family 7 Glycoside Hydrolases

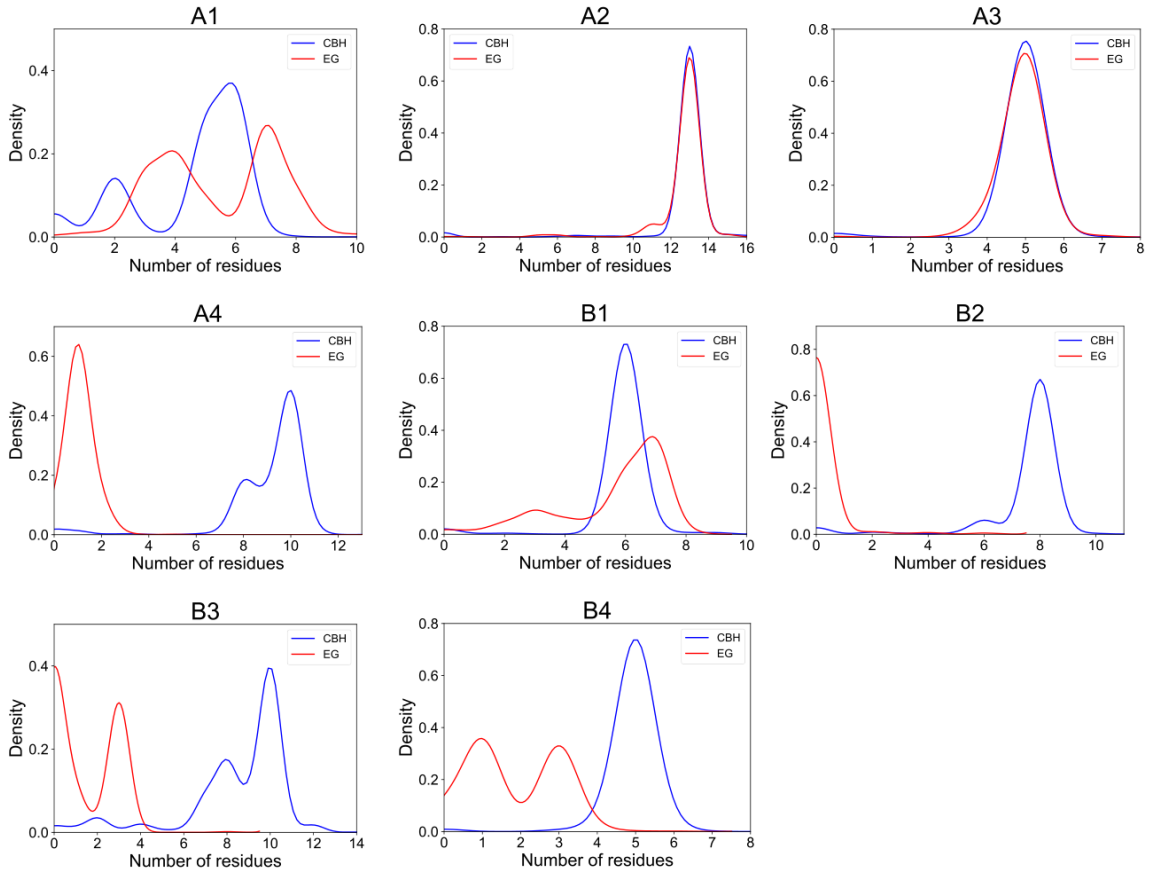
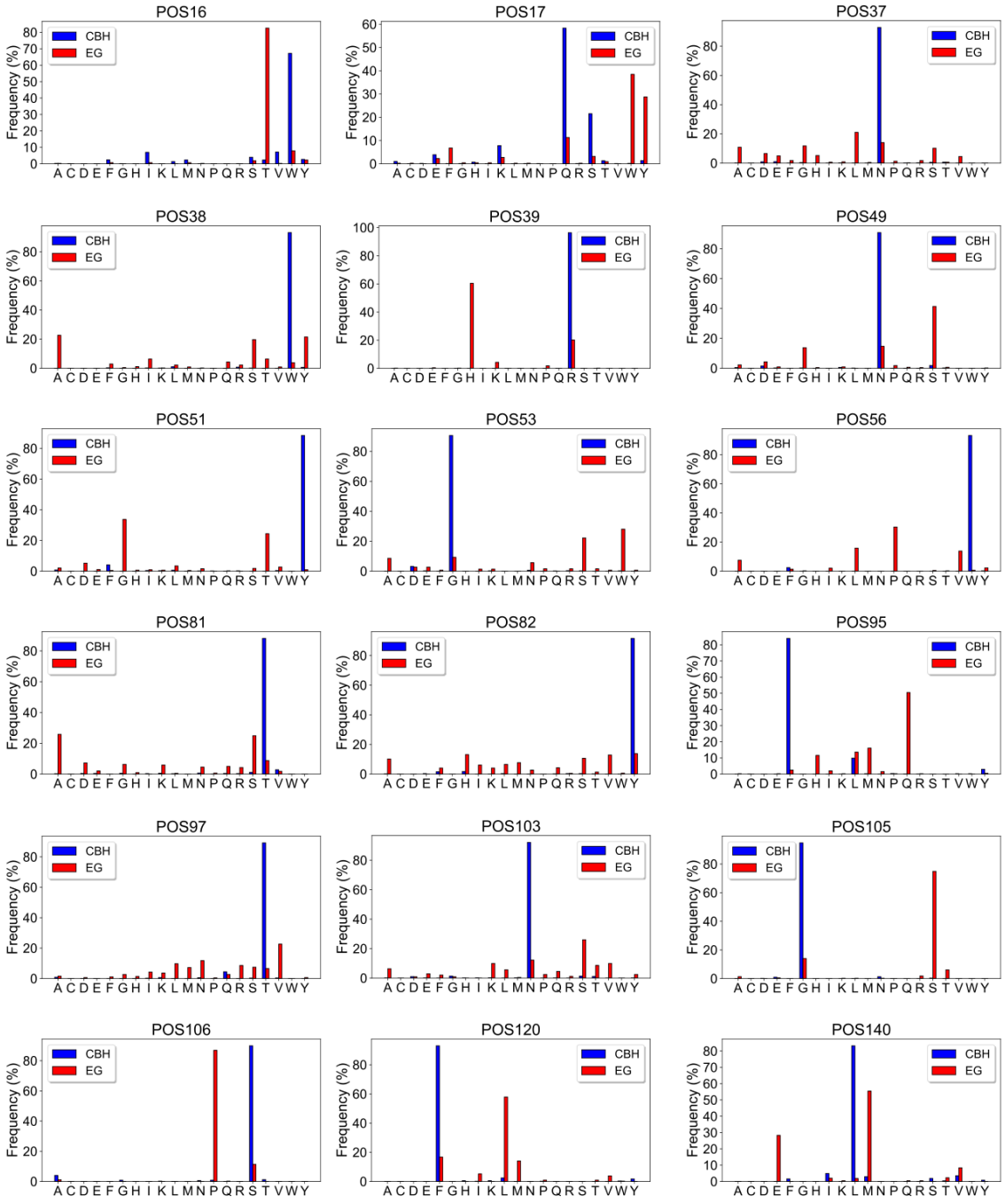
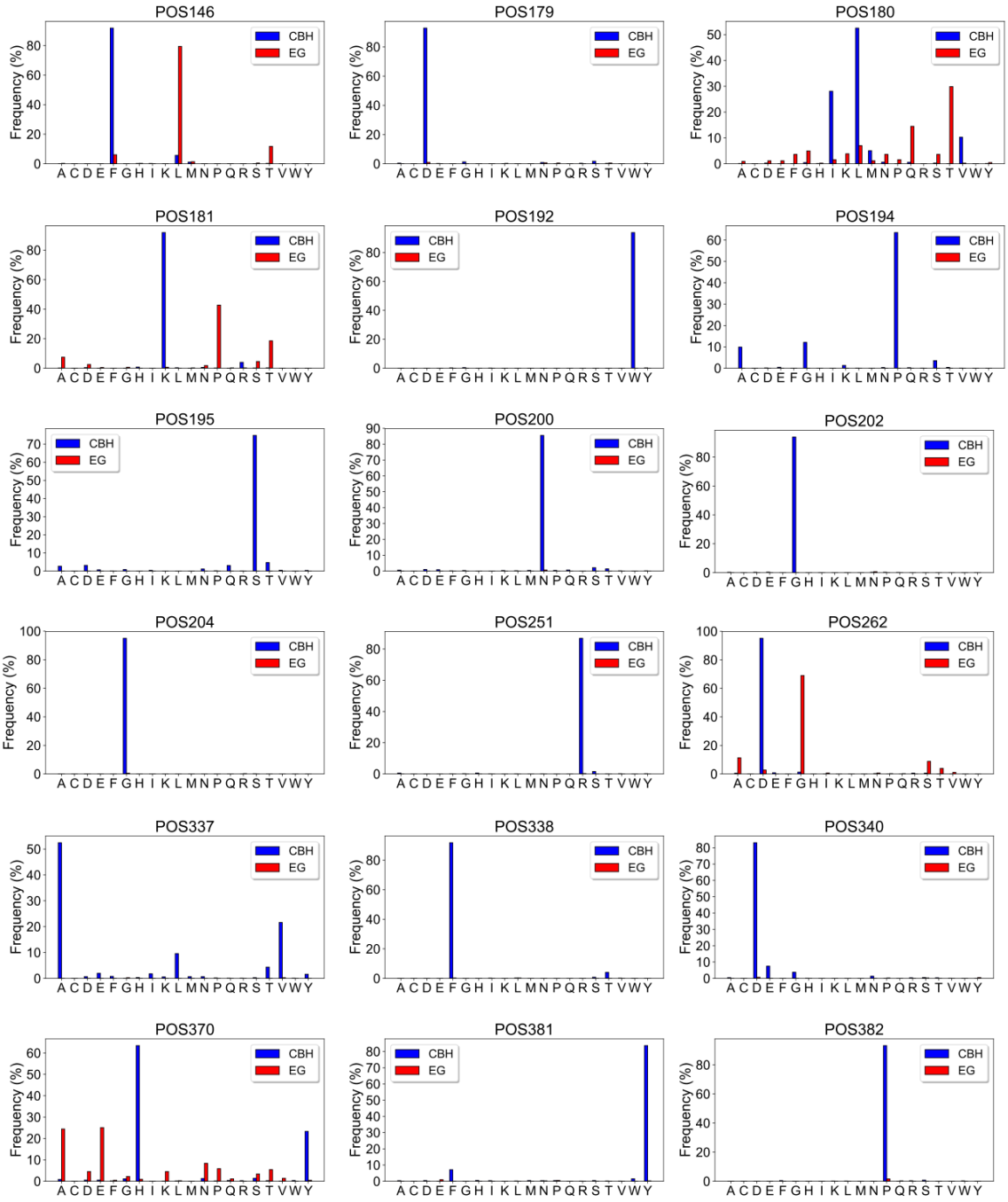


Figure A1.1. Density plots showing the distribution of the lengths of active-site loops in 1,306 GH7 CBHs and 442 EGs. The A2 and A3 loops show nearly identical distributions among CBHs and EGs; the A1 and B1 loops show fairly different distributions; and the A4, B2, B3, and B4 show clearly distinct distributions, with CBHs being notably longer in these four loops.


POS16
POS17
POS37
POS38
POS39
POS49
POS51
POS53
POS56
POS81
POS82
POS95
POS97
POS103
POS105
POS106
POS120
POS140
CBH
EG
Frequency (%)
A C D E F G H I K L M N P Q R S T V W Y


POS146
POS179
POS180
POS181
POS192
POS194
POS195
POS200
POS202
POS204
POS251
POS262
POS337
POS338
POS340
POS370
POS381
POS382
CBH
EG
Frequency (%)
A C D E F G H I K L M N P Q R S T V W Y

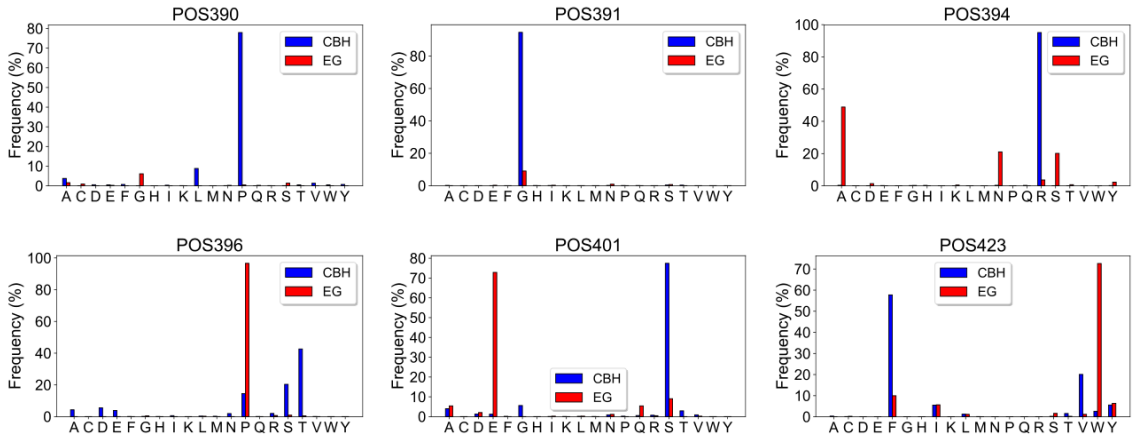


Figure A1.2. Amino acid distribution at 42 positions from which top-performing position-specific classification rules were derived. The frequency of amino acids was determined from structure-based multiple sequence alignments (MSA) of 1,306 CBHs and 442 EGs. Positions in the MSA are referred to using *Tre*Cel7A numbering. Gaps are not included in the analysis, hence the total frequency may not equal 100%.

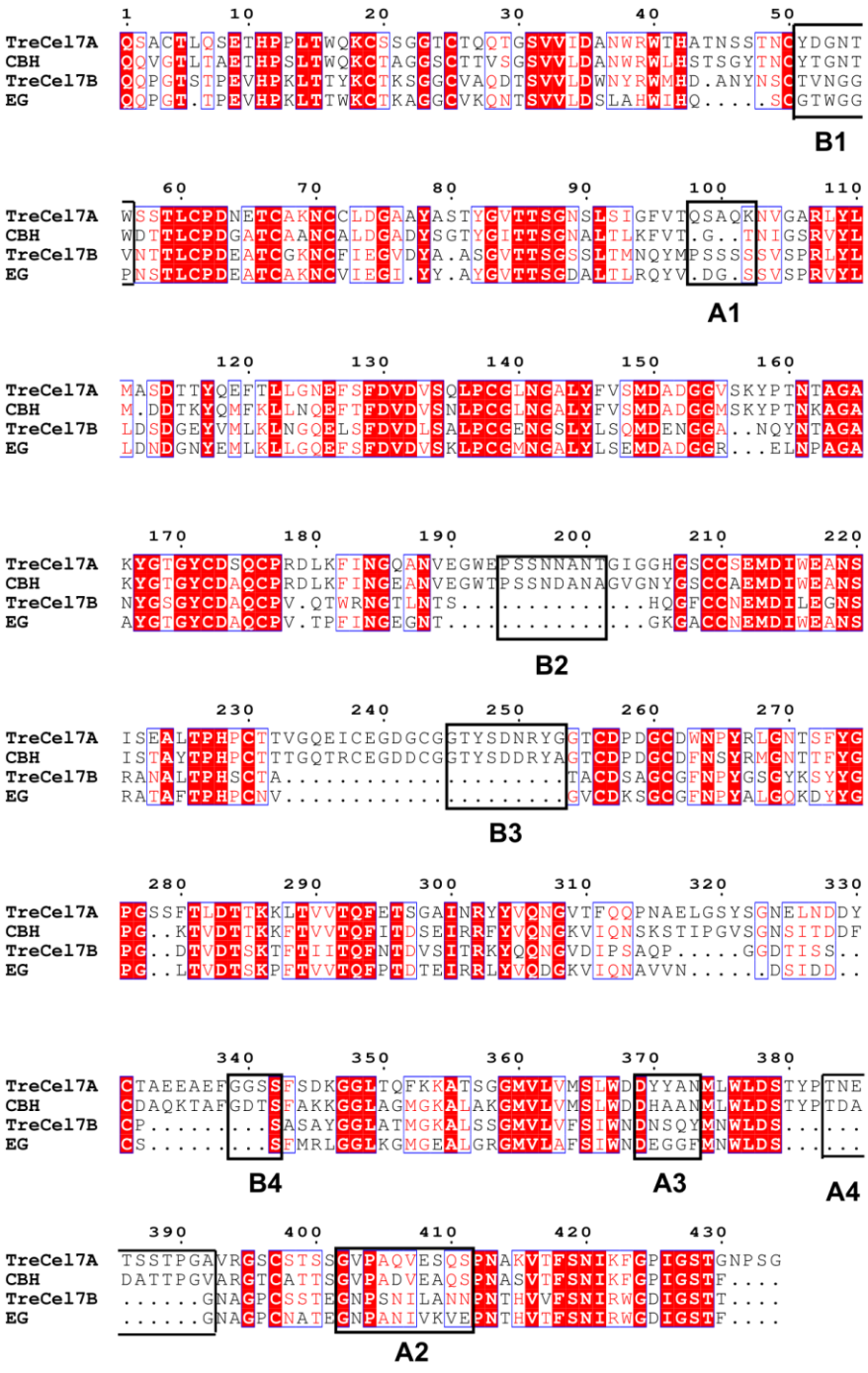
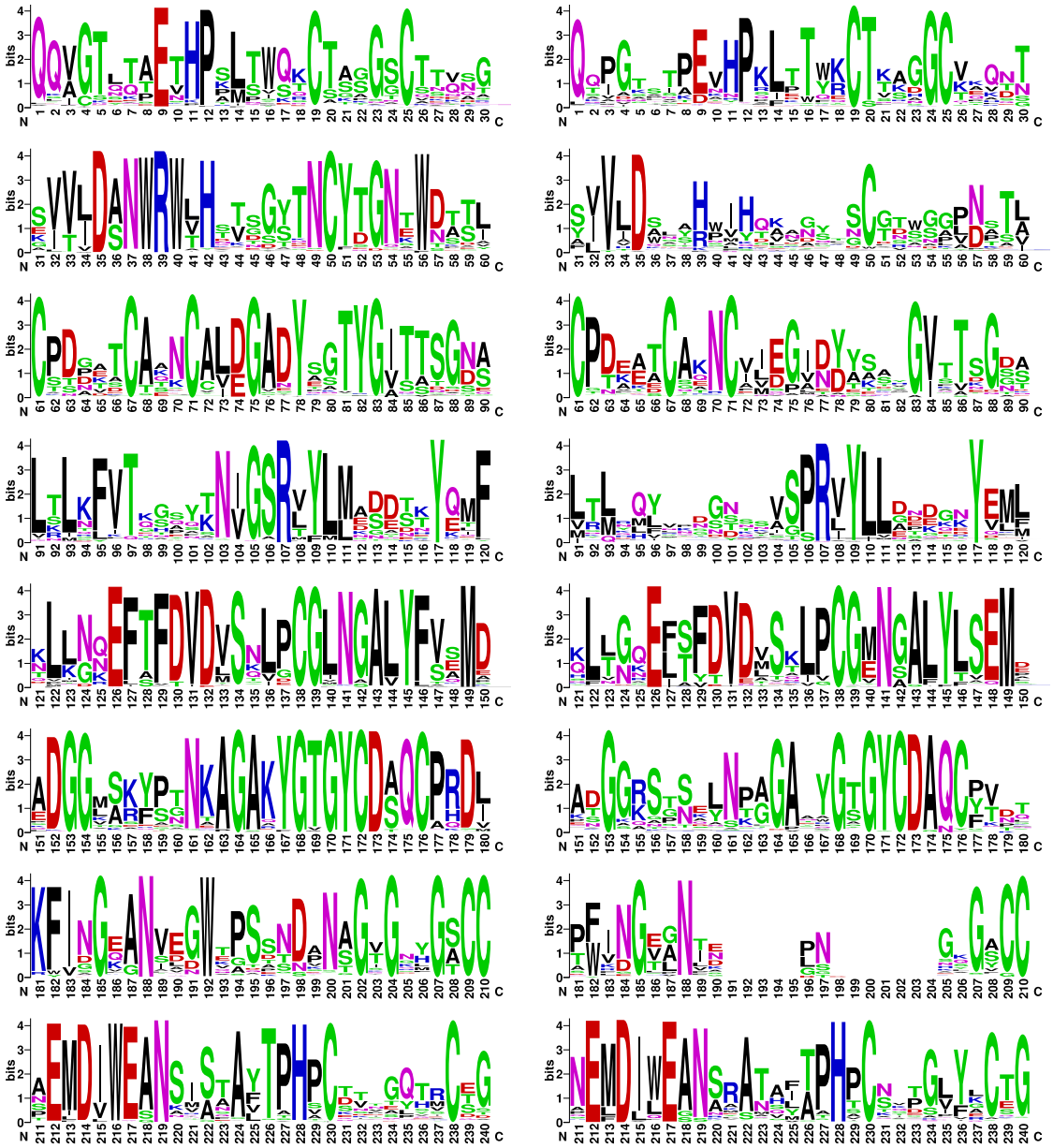


Figure A1.3. Multiple sequence alignment (MSA) of a typical GH7 CBH (*Tre*Cel7A), a typical GH7 EG (*Tre*Cel7B), and the consensus GH7 CBH and EG sequences. The consensus sequences were determined from MSAs of 1,306 GH7 CBHs and 442 GH7 EGs described in the Materials and Methods section of the main text. The eight active-site loop

regions are shown in black boxes. Positions are labelled with *Tre*Cel7A numbering and only positions corresponding to *Tre*Cel7A residues are shown. The alignment was viewed with ESPript (http://espript.ibcp.fr/).



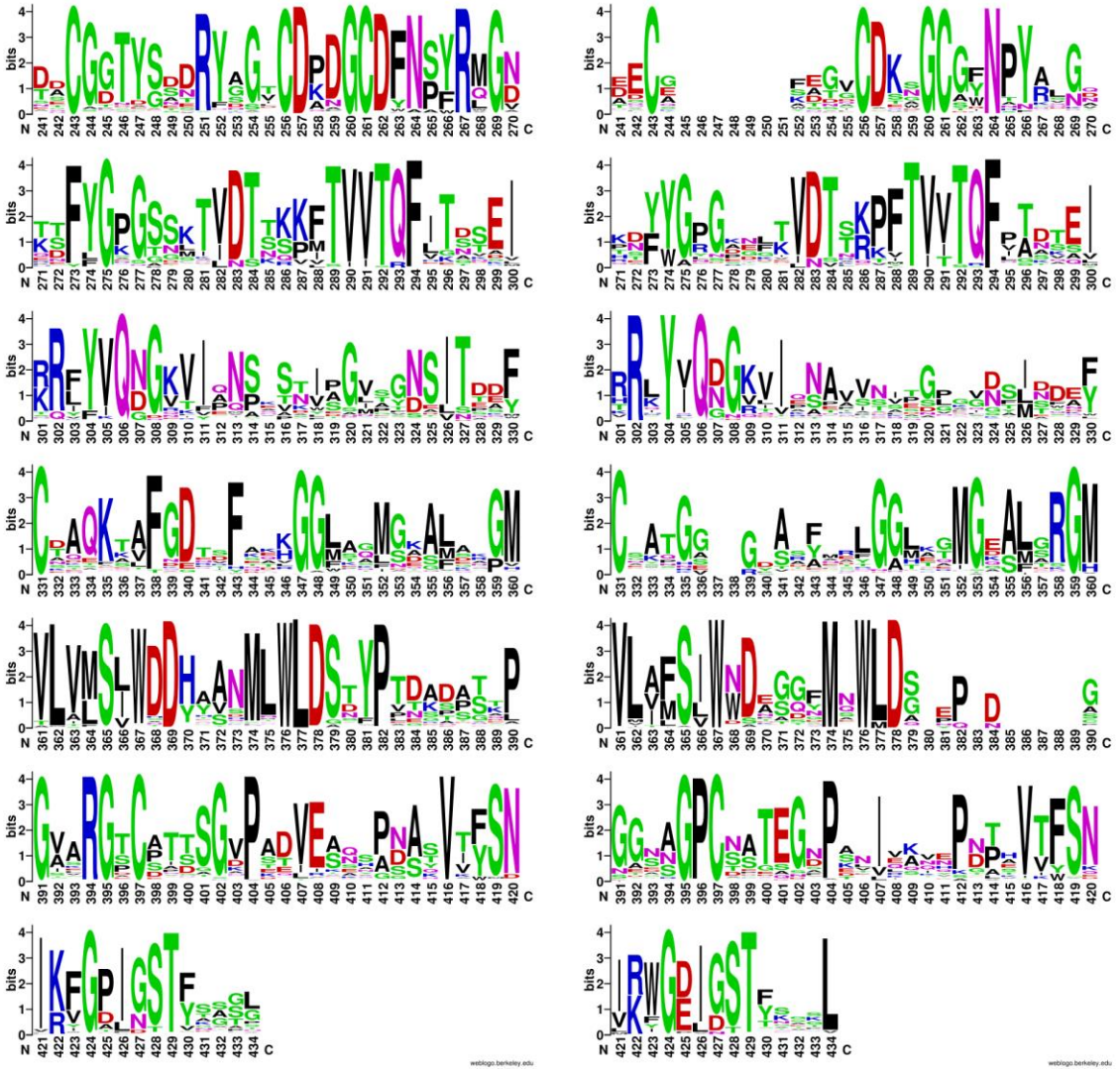


Figure A1.4. Sequence logos of GH7 CBH (left) and EG (right). The logos were derived from the structure-based multiple sequence alignments (1,306 and 442 sequences, respectively). Only positions corresponding to residues in *Tre*Cel7A CD are shown, and they are numbered in the sequence logo as such (i.e., residues 1 to 434).

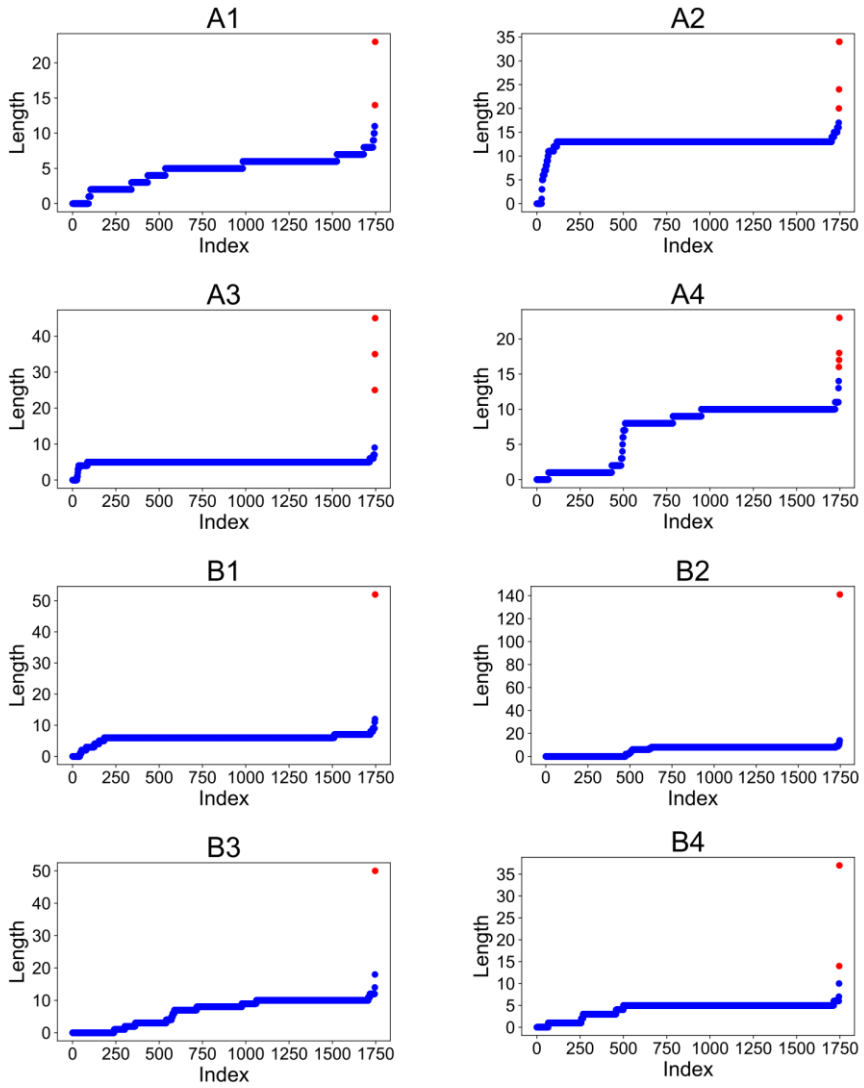
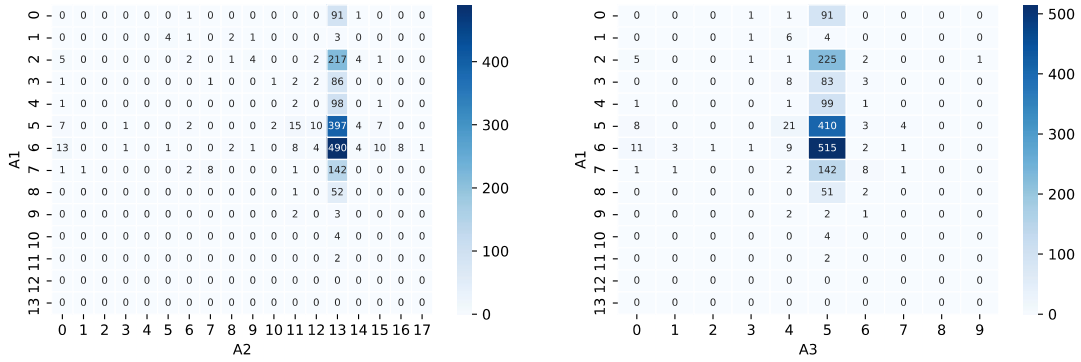
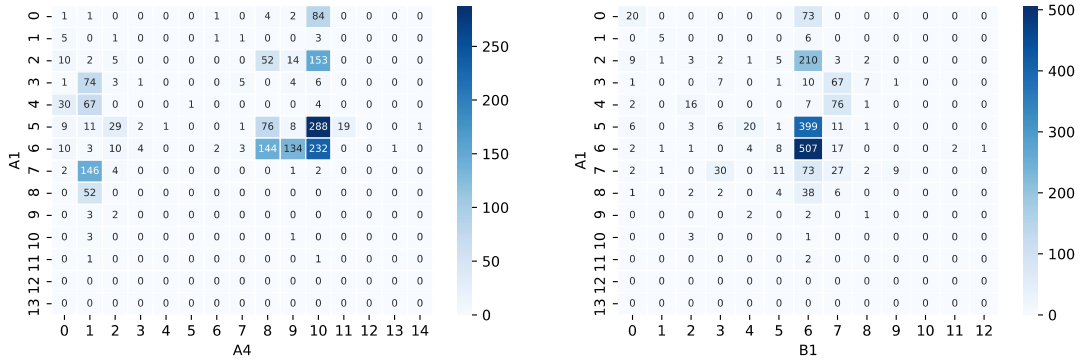
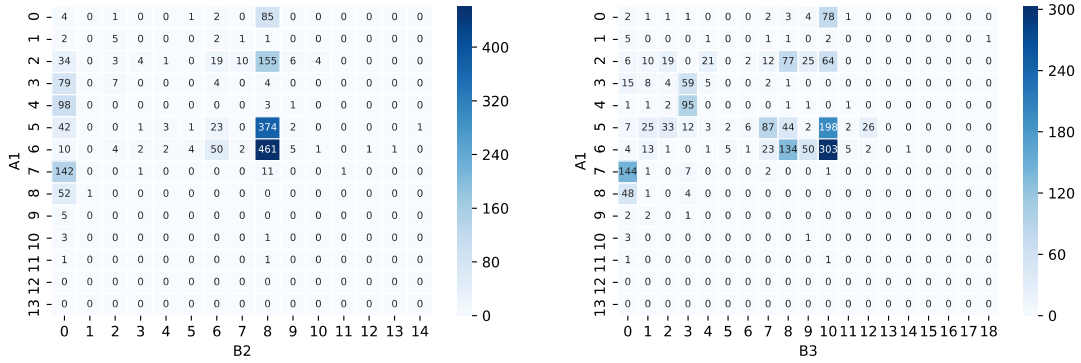
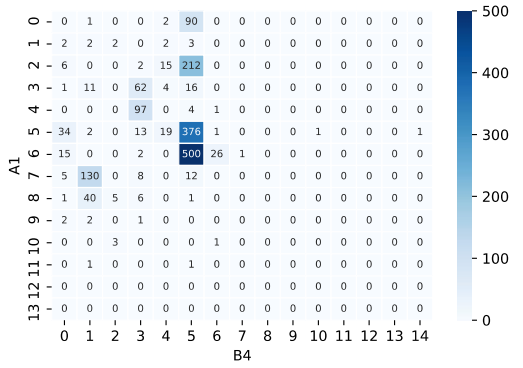


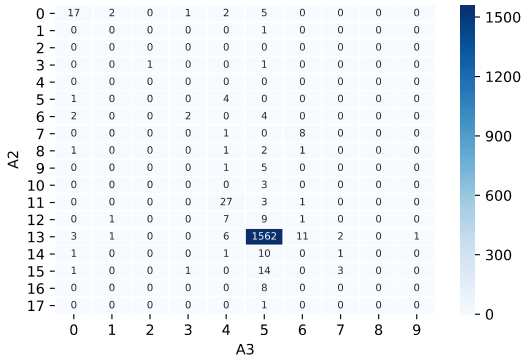
Figure A1.5. Dealing with outliers in active-site loop lengths before machine learning. Each plot shows, in ascending order, the lengths (number of residues) of the active-site loops of 1,748 sequences. Extreme values (points in red) are arbitrarily determined to be outliers. All outliers in the dataset (red points) are capped to a maximum limit (i.e. the maximum length of blue points) before machine learning is applied to the dataset.

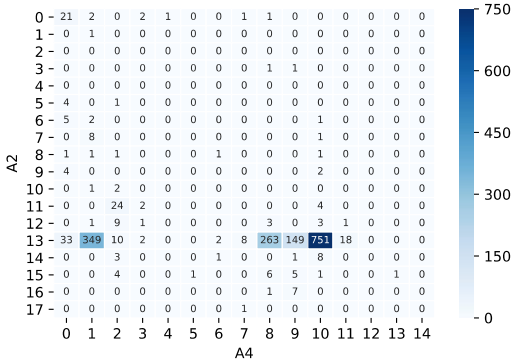

A1
A2
A3

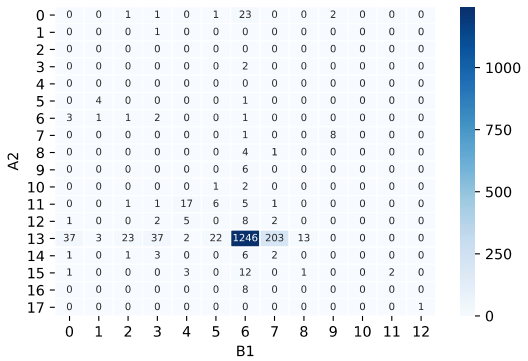

A1
A4
B1

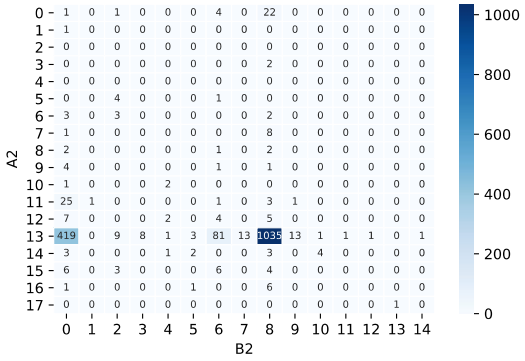

A1
B2
B3

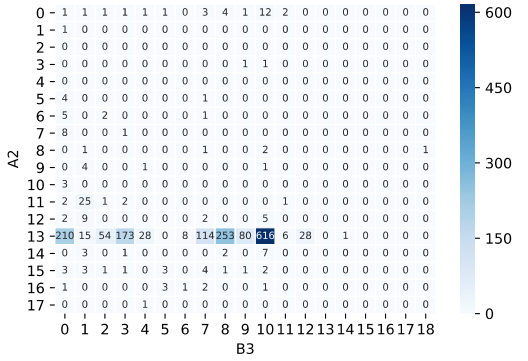

A1
B4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9 10 11 12 13
500
400
300
200
100
0

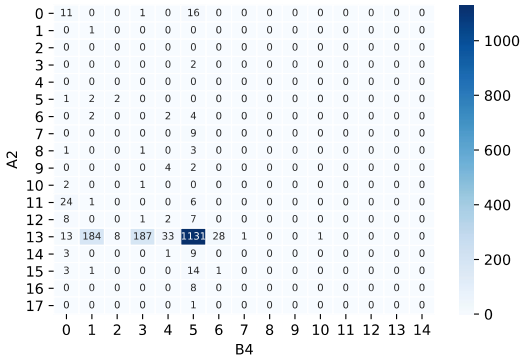

A2
A3
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
1500
1200
900
600
300
0

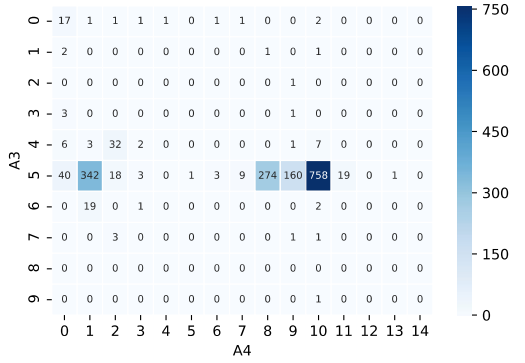

A2
A4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
750
600
450
300
150
0

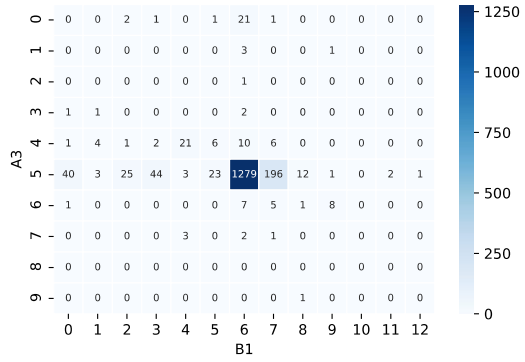

A2
B1
0 1 2 3 4 5 6 7 8 9 10 11 12
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
1000
750
500
250
0

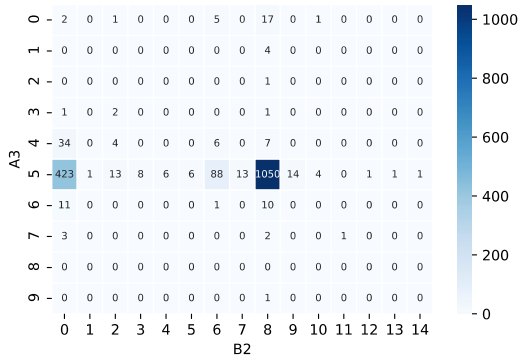

A2
B2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
1000
800
600
400
200
0

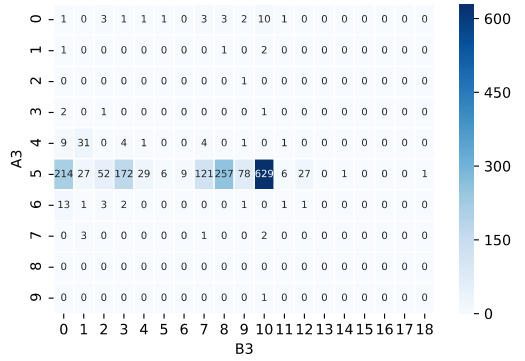

A2
B3
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
600
450
300
150
0

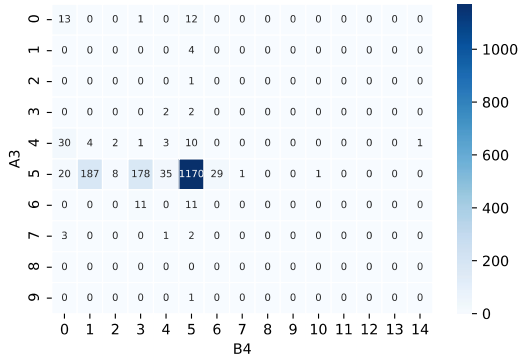

A2
B4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
0 200 400 600 800 1000

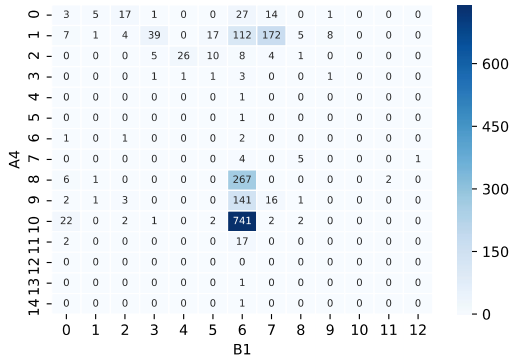

A3
A4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9
0 150 300 450 600 750

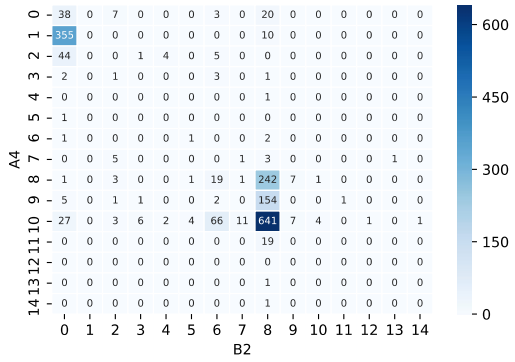

A3
B1
0 1 2 3 4 5 6 7 8 9 10 11 12
0 1 2 3 4 5 6 7 8 9
0 250 500 750 1000 1250

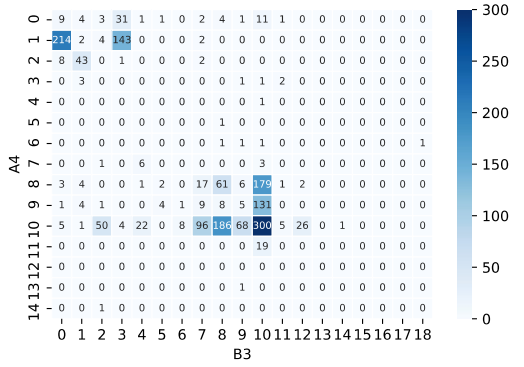

A3
B2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9
0 200 400 600 800 1000

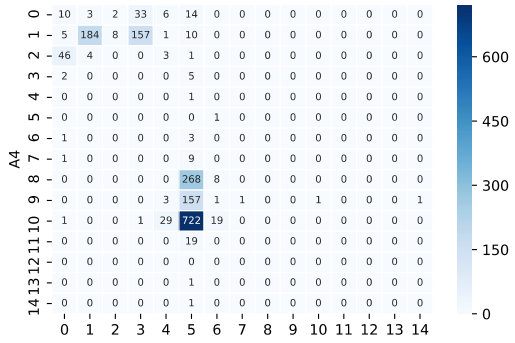

A3
B3
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
0 1 2 3 4 5 6 7 8 9
0 150 300 450 600

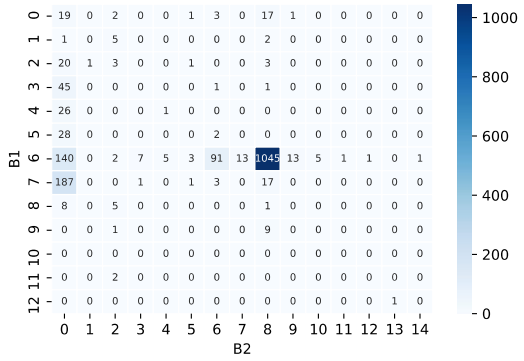

A3
B4
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
0 1 2 3 4 5 6 7 8 9
0 200 400 600 800 1000

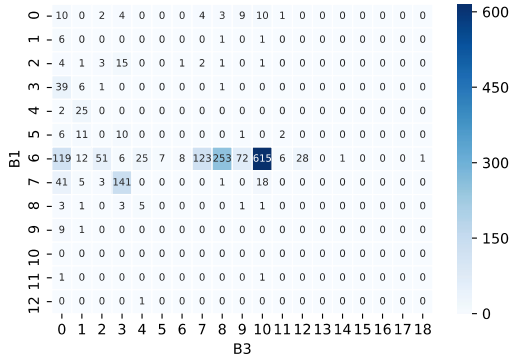

A4
B1


A4
B2


A4
B3


A4
B4


B1
B2


B1
B3

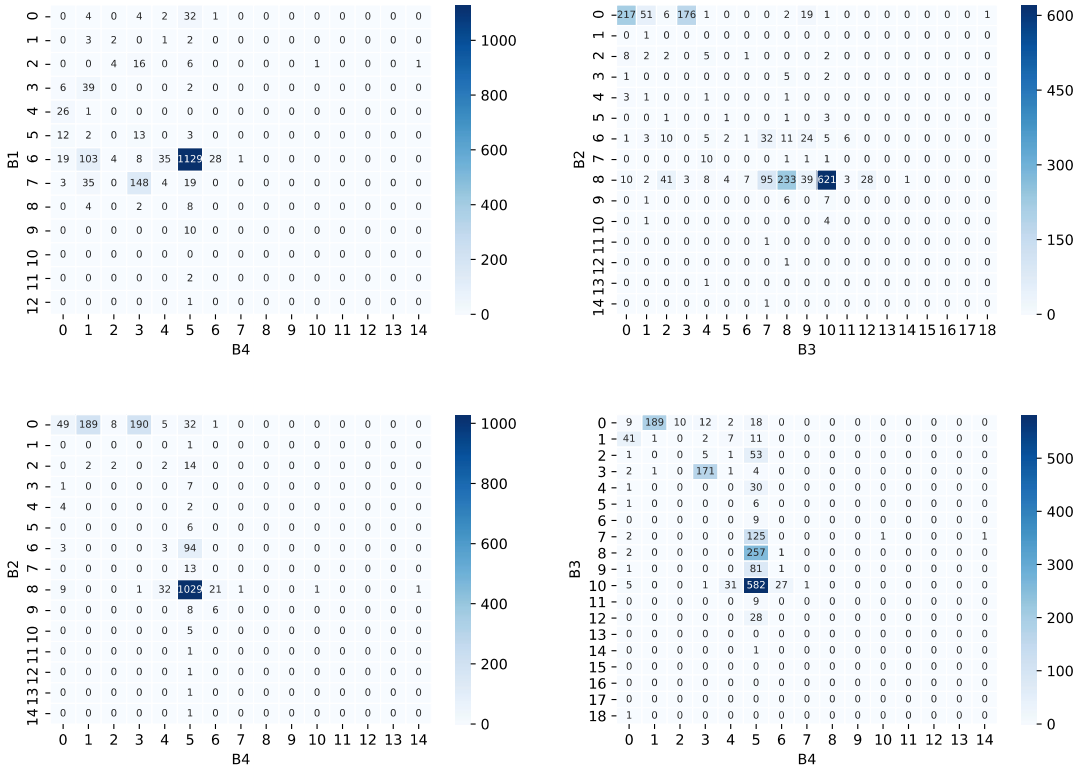


Figure A1.6. Pairwise distribution of the residue lengths of GH7 active-site loops. Each cell in the matrix indicates the number of sequences in the dataset of 1,748 sequences that possess the corresponding number of residues in each loop. For example, there are 490 sequences with six residues in the A1 loop and 13 residues in the A2 loop.

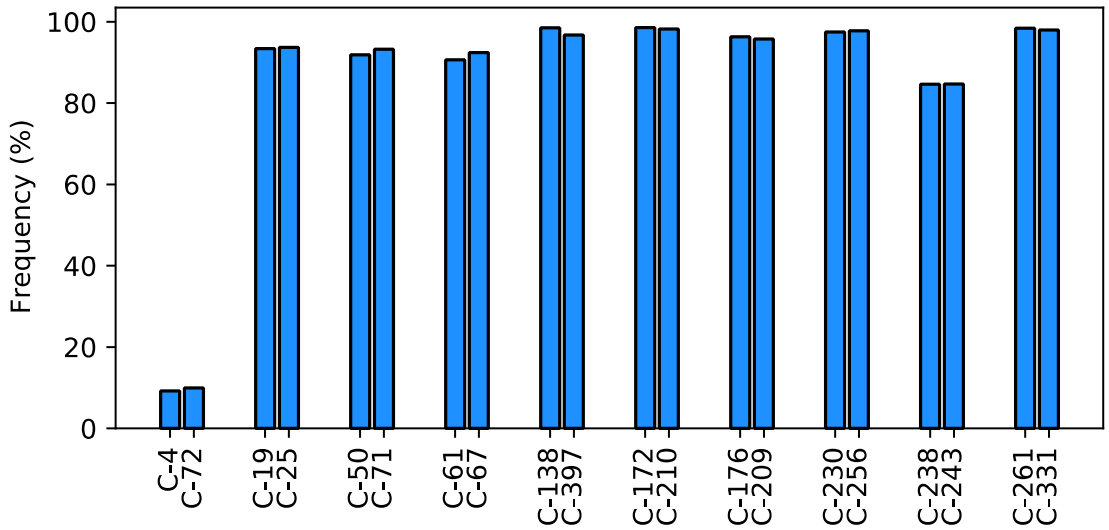


Figure A1.7. Frequency of Cys at positions forming disulfide bonds in GH7 sequences. Cys positions (*x*-axis) are labeled using *Tre*Cel7A numbering and the frequencies were determined from the structure-based multiple sequence alignment of 1,748 sequences. GH7 sequences may have up to ten disulfide bonds, nine of which are present in roughly at least 80% of the sequences. A rare disulfide bond, formed by C4 and C72 in *Tre*Cel7A, is present in less than 10% of GH7 sequences.

Table A1.1. Top-performing position-specific classification rules between amino acid type and GH7 subtype (CBH/EG). Amino acids are grouped into the following types: ALI – aliphatic (Ala, Gly, Val, Leu, Ile, Met, Cys, and Pro), ARO – aromatic (Phe, Trp, Tyr, His), POS – positive (Arg, Lys), NEG – negative (Asp, Glu), and POL – polar (Asp, Gln, Ser, Thr). Positions are represented with *Tre*Cel7A numbering. All rules discriminate GH7 CBHs and EGs with accuracies of at least 87.0% and MCC scores of at least 0.73. Statistical significance was evaluated by a chi-squared test of independence. All rules are significant at $p < 0.0001$.

*Tre*Cel7A position	Rule	Closest subsite	Distance to closest subsite (Å)	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
16	not POL=>CBH, POL=>EG	-2	19.0	93.7	84.6	91.4	0.78
17	not ARO=>CBH, ARO=>EG	-1	21.6	97.7	74.4	91.8	0.78
39	POS=>CBH, not POS=>EG	-5	3.6	96.5	75.6	91.2	0.76
39	POS=>CBH, ARO=>EG	-5	3.6	98.2	68.0	90.6	0.74
51	ARO=>CBH, not ARO=>EG	-5	3.6	92.6	98.0	94.0	0.86
56	ARO=>CBH, not ARO=>EG	-5	8.9	95.9	95.7	95.8	0.89
56	not ALI=>CBH, ALI=>EG	-5	8.9	100.0	69.7	92.3	0.79
56	ARO=>CBH, ALI=>EG	-5	8.9	97.9	82.7	94.1	0.84
105	ALI=>CBH, not ALI=>EG	-4	4.8	95.0	84.4	92.3	0.80
105	not POL=>CBH, POL=>EG	-4	4.8	98.4	81.2	94.1	0.84
105	ALI=>CBH, POL=>EG	-4	4.8	96.7	82.8	93.2	0.82
106	POL=>CBH, not POL=>EG	-2	4.8	91.7	88.5	90.9	0.77
106	not ALI=>CBH, ALI=>EG	-2	4.8	94.3	88.0	92.7	0.81
106	POL=>CBH, ALI=>EG	-2	4.8	93.0	88.2	91.8	0.79
120	ARO=>CBH, not ARO=>EG	-1	15.7	95.6	83.0	92.4	0.80
120	not ALI=>CBH, ALI=>EG	-1	15.7	97.2	81.9	93.3	0.82
120	ARO=>CBH, ALI=>EG	-1	15.7	96.4	82.5	92.9	0.81
146	ARO=>CBH, not ARO=>EG	-1	7.9	92.0	93.7	92.4	0.81
146	not ALI=>CBH, ALI=>EG	-1	7.9	93.0	81.0	90.0	0.74
146	ARO=>CBH, ALI=>EG	-1	7.9	92.5	87.3	91.2	0.78
179	NEG=>CBH, not NEG=>EG	-3	2.6	93.0	99.1	94.6	0.87
180	ALI=>CBH, not ALI=>EG	-3	4.1	96.5	82.6	93.0	0.81
181	POS=>CBH, not POS=>EG	-5	2.8	95.9	99.1	96.7	0.92
181	POS=>CBH, ALI=>EG	-5	2.8	97.6	75.1	91.9	0.78

Table A1.1 (continued)

*Tre*Cel7A position	Rule	Closest subsite	Distance to closest subsite (Å)	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
192	ARO=>CBH, not ARO=>EG	-4	7.0	94.3	100.0	95.7	0.90
194	ALI=>CBH, not ALI=>EG	-4	11.0	85.7	100.0	89.3	0.78
195	POL=>CBH, not POL=>EG	-3	8.1	83.7	100.0	87.8	0.75
200	POL=>CBH, not POL=>EG	-4	3.5	89.7	99.3	92.2	0.82
202	ALI=>CBH, not ALI=>EG	-4	6.5	94.6	100.0	95.9	0.90
204	ALI=>CBH, not ALI=>EG	-4	10.7	95.2	99.5	96.3	0.91
251	POS=>CBH, not POS=>EG	2	3.4	86.9	99.8	90.2	0.79
262	NEG=>CBH, not NEG=>EG	2	4.1	95.9	97.3	96.3	0.91
262	not ALI=>CBH, ALI=>EG	2	4.1	98.2	82.4	94.2	0.84
262	NEG=>CBH, ALI=>EG	2	4.1	97.1	89.8	95.2	0.87
337	ALI=>CBH, not ALI=>EG	2	11.6	86.2	99.5	89.6	0.78
338	ARO=>CBH, not ARO=>EG	2	7.7	92.0	99.8	94.0	0.86
340	NEG=>CBH, not NEG=>EG	2	9.1	90.6	99.3	92.8	0.84
370	ARO=>CBH, not ARO=>EG	-3	5.3	87.3	98.2	90.0	0.78
381	ARO=>CBH, not ARO=>EG	2	3.5	92.8	99.8	94.6	0.87
382	ALI=>CBH, not ALI=>EG	2	5.0	93.5	98.4	94.7	0.87
390	ALI=>CBH, not ALI=>EG	2	8.8	92.1	91.0	91.8	0.80
391	ALI=>CBH, not ALI=>EG	2	6.9	94.9	90.7	93.9	0.84
394	POS=>CBH, not POS=>EG	2	3.1	95.1	95.9	95.3	0.88
394	POS=>CBH, ALI=>EG	2	3.1	97.3	72.5	91.0	0.75
401	not NEG=>CBH, NEG=>EG	-3	13.5	97.5	74.9	91.8	0.77

Table A1.2. Top-performing position-specific classification rules for predicting the presence of CBMs in GH7s. The rules were derived from the validation dataset (1,574 sequences). The sensitivity and specificity indicate the percent of GH7s with CBMs and the percent of GH7s without CBMs correctly classified by each rule, respectively. The top 5 rules (ranked by MCC scores) are shown in bold, and are derived from residues at the C-terminus, or residues consisting the tenth rare disulfide bond found in *Tre*Cel7A. Statistical significance was evaluated by a chi-squared test of independence. All rules are significant at $p < 0.0001$.

Position	Rule	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
2	Q2=>CBM	83.3	44.9	55.3	0.17
4	**C4=>CBM**	**22.5**	**95.6**	**75.8**	**0.47**
22	S22=>CBM	43.4	79.7	69.9	0.23
37	N37=>CBM	90.8	34.1	49.4	0.15
40	W40=>CBM	95.8	31.2	48.7	0.17
42	H42=>CBM	96.9	25.1	44.5	0.14
44	T44=>CBM	52.3	73.3	67.6	0.22
48	T48=>CBM	81.2	49.3	57.9	0.19
66	T66=>CBM	74.6	50.6	57.1	0.16
72	**C72=>CBM**	**23.0**	**94.7**	**75.3**	**0.43**
76	A76=>CBM	88.0	40.1	53.0	0.17
87	S87=>CBM	88.3	43.5	55.6	0.19
94	N94=>CBM	30.8	90.1	74.0	0.34
103	N103=>CBM	91.5	36.1	51.1	0.17
123	L123=>CBM	75.8	53.1	59.3	0.18
135	N135=>CBM	53.8	72.7	67.6	0.22
135	not K135=>CBM	94.1	28.1	46.0	0.14
140	not M140=>CBM	98.6	21.3	42.2	0.14
174	S174=>CBM	57.5	74.3	69.8	0.26
177	P177=>CBM	92.5	39.9	54.1	0.19
183	not V183=>CBM	99.8	17.2	39.6	0.13
221	I221=>CBM	54.2	72.2	67.3	0.22
226	T226=>CBM	98.8	21.0	42.1	0.14
229	P229=>CBM	83.8	42.8	53.9	0.16
258	not K258=>CBM	80.0	43.3	53.2	0.14
268	Q268=>CBM	32.2	88.9	73.6	0.32

Table A1.2 (continued)

Position	Rule	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
272	not D272=>CBM	93.4	26.6	44.7	0.13
296	T296=>CBM	81.2	48.3	57.2	0.18
296	not G296=>CBM	100.0	10.3	34.6	0.09
307	N307=>CBM	79.1	60.5	65.5	0.26
336	T336=>CBM	45.8	80.6	71.2	0.27
341	T341=>CBM	44.4	78.8	69.5	0.23
364	L364=>CBM	38.7	80.8	69.4	0.21
371	A371=>CBM	49.3	79.3	71.2	0.27
375	L375=>CBM	81.5	47.3	56.5	0.17
379	S379=>CBM	95.8	33.3	50.2	0.18
396	not P396=>CBM	83.1	41.9	53.0	0.15
398	S398=>CBM	38.0	85.3	72.5	0.29
409	A409=>CBM	53.8	73.1	67.9	0.22
410	N410=>CBM	41.1	87.6	75.0	0.37
413	not D413=>CBM	88.0	36.0	50.1	0.14
413	N413=>CBM	61.5	69.9	67.7	0.24
414	S414=>CBM	36.9	87.2	73.6	0.32
421	I421=>CBM	98.4	20.2	41.4	0.13
431	S431=>CBM	39.2	84.3	72.1	0.28
432	**G432=>CBM**	**35.0**	**93.2**	**77.4**	**0.49**
433	S433=>CBM	24.4	91.5	73.3	0.31
433	**G433=>CBM**	**32.6**	**95.4**	**78.4**	**0.57**
433	**T433=>CBM**	**17.8**	**96.8**	**75.4**	**0.49**
434	**L434=>CBM**	**28.4**	**96.6**	**78.1**	**0.60**

A2 Supporting Information for Enabling Microbial Syringol Conversion Through Structure-Guided Protein Engineering

A2.1 Supplementary Materials and Methods

A2.1.1 Protein expression and purification

Expression constructs were expressed in E. coli Rosetta™ 2 (DE3) cells (Novagen). Cells were transformed with plasmids for expression of the GcoA mutants (pGcoA-F196A, pGcoA-F196H, pGcoA-F196I, pGcoA-F196L, pGcoA-F196S, or pGcoA-F196V) and plated out lysogeny broth (LB) agar containing chloramphenicol (34 mg/L) and carbenicillin (50 mg/L). A single colony was selected and used to inoculate a 20 mL starter culture of LB. After overnight growth at 37 ºC, 250 rpm, the starter culture was inoculated into 2.5 L flasks containing 1 L of terrific broth (TB) with antibiotics. At an OD600 of 1.0, 0.2 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to induce protein expression. 100 mg/L 5-aminolevulinic acid (GcoA) was added to support productive cofactor incorporation. Induction of protein expression was performed for 16-18 hr at 20 ºC with shaking at 250 rpm. Affinity purification was carried out using glutathione-sepharose 4B media (GE Lifesciences) followed by GST-tag cleavage with PreScission protease (GE Lifesciences). Anion exchange chromatography was performed with Source 30Q media (GE Lifesciences) packed into a (GE HR 16/100 Column) with a 10-40% gradient of buffers A (50 mM HEPES pH 7.5, 100 mM NaCl, 1 mM DTT) and B (50 mM HEPES pH 7.5, 1 M NaCl, 1 mM DTT) respectively. For each protein, a final gel filtration

step was performed using a HiLoad S200 16/60 pg column (GE Lifesciences) in a buffer containing 25 mM HEPES pH 7.5 and 50 mM NaCl.

A2.1.2 Cofactor Analyses.

Heme Quantification. To determine the amount of catalytically active heme, CO gas was bubbled into a cuvette containing 1.0-2.5 μM (Pierce BCA assay) F169 GcoA mutants (A,H,S,V,I,L), made up in buffer (25 mM HEPES, 50 mM NaCl, pH 7.5) containing 1.0 mM EDTA, 20% glycerol, 0.5% sodium cholate, and 0.4% non-ionic detergent. Excess sodium dithionite (~1 mg) was added to reduce the heme iron and the peak attributed to the catalytically competent, ferrous CO-bound heme (~450 nm) gradually appeared. Several scans were taken to ensure complete binding of CO to heme. A spectrum for a control containing only dithionite-reduced GcoA was measured, and a difference spectrum computed. Absorbances at 420, 450, and 490 nm were recorded to calculate the amount of active GcoA (P450) or inactive GcoA (P420 nm) (see Equations A2.1 to A2.3). Reported values are the average of three or more measurements.

$$(\Delta A450 - \Delta A490)/0.091 = \text{nmol of P450 per mL} \quad \text{(A2.1)}$$

$$[(\Delta A420 - A490)_{observed} - (A450 - A490)_{theoretical}]/0.110 = \text{nmol of cytochrome P420 per mL} \quad \text{(A2.2)}$$

$$\text{nmol of P450 per mL} \times (-0.041) = (\Delta A420 - A490)_{theoretical} \quad \text{(A2.3)}$$

Here ΔA450 and ΔA420 are the differences between the reference and sample spectra at absorbances 450 and 420 nm, respectively.

Determination of [FAD] and non-heme [Fe] in GcoB. FAD was released from GcoB by denaturing 200 µL of a protein (0.024 µM) solution with 5 µL saturated ammonium sulfate, pH 1.4 (7% v/v H_2SO_4). Precipitated protein was pelleted by centrifugation and the UV/vis spectrum of the FAD-containing supernatant was measured. The absorbance at 454 nm, ε_{FAD} = 11.3 mM^{-1} cm^{-1}, and total protein concentration determined by the BCA assay (Pierce) were used to determine [FAD] bound to GcoB. An extinction coefficient for GcoB-bound FAD was estimated via the slope of a line relating absorbance at 454 nm to [GcoB-FAD].

To determine the Fe-S content of GcoB, the protein was first denatured as described above. 50 µL of supernatant was added to 25 µL of 5% w/v sodium ascorbate to reduce the iron. 100 µL of bathophenanthroline disulfonate (0.1% w/v in ddH_2O) was added and the sample was incubated for 1h. The resulting Fe(II) complex was quantified via its absorbance at 535 nm (ε_{535} = 22.14 mM^{-1} cm^{-1}, determined using $FeSO_4$ standards). An extinction coefficient for GcoB-bound 2Fe-2S cluster was estimated via the slope of a line relating absorbance at 423 nm to [GcoB-2Fe-2S].

A2.1.3 Steady state kinetics and substrate dissociation constants.

Steady state kinetics of F169A. 0.2 µM each of F169A GcoA and GcoB were dissolved in air-saturated buffer (25 mM HEPES, 50 mM NaCl) in a cuvette at pH 7.5, 25 °C. 100 µg/mL catalase was added to each reaction to capture any H_2O_2 formed during the uncoupled reaction. A saturating amount of NADH (≥ $5K_M$, 300 µM) was added and a background rate of NADH oxidation in air (~210 µM O_2) recorded via continuous scanning

of the UV/vis spectrum (Varian Cary 50). 20-300 µM guaiacol or syringol 2-20 mM stock dissolved in DMSO was added and the reaction was monitored via measurement of UV/vis spectra for several minutes. The initial velocity was determined by disappearance of the characteristic NADH absorbance at 340 nm (ε_{344} = 6.22 mM^{-1} cm^{-1}). A plot of v_i versus [guaiacol] was fit to Equation A2.4 to obtain k_{cat}, K_M, *and* k_{cat}/K_M. For specific activity determination, the above method was used but with saturating (300 µM) guaiacol, syringol, or 3-methoxycatechol (3MC), and in the presence of all F169 GcoA mutants (A,H,S,V,I,L). The linear portion of [NADH] vs time was fit and referenced to the amount of GcoA used (0.2 µM). Reported values are the average of ≥3 measurements and reported errors are standard deviations.

$$v_i = V_{max}[S]/(K_M + [S]) \qquad \text{(A2.4)}$$

Determination of substrate dissociation constants (K_D) with F169A. 0-60 µM of guaiacol, syringol, or 3MC in 0.5 or 1 µM aliquots, were titrated into a cuvette containing 3 µM F169A GcoA in 25 mM HEPES, 50 mM NaCl, pH 7.5. The spectrum after each substrate addition was recorded, beginning with no substrate bound. The solution reached equilibrium before the next addition. A difference spectrum was made to illustrate the shift from a low-spin aquo-heme complex to the high-spin substrate-bound complex (spectral shift from 417 nm to 388 nm). The resulting difference spectra showed a peak at 388 nm, and a trough at 417 nm. The absorbance at 388 nm ($Abs_{388\text{-}417\ nm}$) was plotted as a function of [substrate], yielding a quadratic curve that was fit to Equation A2.5 to determine the K_D.

$$\Delta\, Abs_{obs} = \frac{\Delta\, Abs_{max}}{2E_t}(L_0 \;+\; E_t \;+\; K_D \;- \sqrt{(L_0 \;+\; E_t \;+\; K_d)^2 - 4E_t * L_0)} \tag{A2.5}$$

Where L_0, E_t, K_D, and ΔAbs_{max} are the ligand concentrations, total protein (subunit) concentration, the equilibrium dissociation constant, and the maximum $Abs_{388\text{-}417\ nm}$, respectively. Reported values are the average of 2 or more measurements.

A2.1.4 Product analysis.

Formaldehyde determination. A colorimetric assay using tryptophan can be used to quantify the amount of formaldehyde produced in F169 GcoA/B reactions with guaiacol, syringol, or 3MC. 0.2 µM each of F169 GcoA mutants and GcoB were dissolved in air-saturated buffer (25 mM HEPES, 50 mM NaCl) in a cuvette at pH 7.5, 25 °C. 100 µg/mL catalase was added to each reaction to capture any H_2O_2 formed during the uncoupled reaction. 200 µM NADH was added and the background rate recorded. 100 (guaiacol, syringol, or 3MC) or 200 (syringol) µM of substrate was then added and the reaction monitored until there was no more change, due to either substrate, NADH or O_2 depletion, whichever occurred first. 200 µL of the reaction was then quenched by adding 200 µL of a 0.1% tryptophan solution in 50% ethanol and 200 µL of 90% sulfuric acid. Upon thorough mixing, 40 µL of 1% $FeCl_3$ was added. The solution was then incubated in a heating block for 90 min at 70 °C. After cooling, the absorbance was read at 575 nm and the [formaldehyde] calculated by using $\varepsilon_{575\ nm}$ = 4.2 mM^{-1} cm^{-1}, obtained with formaldehyde as a standard.[70] A negative control included everything but the substrate and was used as a baseline.

HPLC for product identification and specific activity measurement. Analyte analysis of the above end-point reactions (100 μM guaiacol, syringol, or 3MC, or 200 μM syringol) was performed on an Agilent 1100 LC system (Agilent Technologies) equipped with a G1315B diode array detector (DAD). Each sample and standard was injected at a volume of 10 μL onto a Symmetry C18 column 5 μm, 4.6 x 150 mm column (Waters). The column temperature was maintained at 30 °C and the buffers used to separate the analytes of interest was 0.01% TFA in water (A)/ acetonitrile (B). The separation was carried out using a gradient program of: (A) = 99% and (B) = 1% at time t = 0 min; (A) = 99% and (B) = 1% at time t = 2 min, (A) = 50% and (B) = 50% at t = 8 min; (A) = 1% and (B) = 99% at t = 8.01 min; (A) = 99% and (B) = 1% at t = 10.01 min; (A) = 99% and (B) = 1% at t = 11 min. The flow rate was held constant at 1.5 mL min^{-1}, resulting in a run time of 11 minutes. DAD wavelengths of 210 and 325 nm were used for analysis of the analytes of interest. Standard curves were generated using 0-500 μM of guaiacol, syringol, 3MC, catechol, and pyrogallol. Integrated intensities vs [standards] were plotted and the resulting standard curves used to quantify the reactants and products.

For specific activity determination, 300 μM guaiacol, syringol, or 3MC were added from 0.1 M DMSO stocks to air saturated buffer (25 mM HEPES, 50 mM NaCl, pH 7.5), with a final volume of 1 mL. The [analyte] was measured via the above HPLC method. Then, 0.2 μM F169A/ GcoB and 100 μg/mL catalase were added. Upon addition of 300 μM NADH, the timer was started and 50 μL removed every 10 (guaiacol and syringol) or 30 (3MC) seconds. The reaction of each aliquot was immediately quenched with 12.5 μL saturated ammonium sulfate, 7% v/v H_2SO_4 (pH 2.0) prior to loading onto the HPLC

column. [Substrate] disappearance was referenced to GcoA (0.2 μM) and fit to a linear line to determine specific activity.

Data analysis. The consumption of NADH and subsequent production of formaldehyde and aromatic product were measured in triplicate and the error reported as ±1 standard deviation. To determine the statistical significance between NADH consumption and the products produced, the p-value was determined for all runs containing guaiacol, syringol, and 3MC. These values were calculated using two degrees of freedom, a tail value of two, and assuming a t-statistical value for unequal variances.

A2.1.5 Uncoupling reactions.

Detection of H_2O_2 via horseradish peroxidase (HRP) and Amplex Red assay. The reaction between 100 μM guaiacol or 3MC with 0.2 μM F169A/GcoB and 100 μM NADH in air-saturated buffer (25 mM HEPES, 50 mM NaCl, pH 7.5) was monitored continuously in a quartz cuvette, using the NADH consumption assay described above. The same thing was done for syringol, but with either 100 or 200 μM NADH. When there was no longer any change in the spectra, e.g., the reaction was completed, 100 μL was removed from the cuvette and pipetted into a 96-well microplate. A 5 mL solution containing 50 μL of 10 mM Amplex Red (prepared in DMSO and stored at -20 °C) and 100 μL of 10 U/mL HRP was made up in the above buffer. 100 μL of this was added to each of the wells with each reaction sample. The plate was incubated in the dark at room temperature for 30 min, at which point the absorbance at 572 nm was recorded by a Varioskan Lux microplate reader (Thermo Scientific). The absorbance was compared to a standard curve with 0-100 μM H_2O_2 to quantify the amount of peroxide produced in the reactions.

A2.1.6 Crystallography

Purified protein was buffer exchanged into 10 mM HEPES pH 7.5 and concentrated to an A280 value of 12, as measured on a NanoDrop 2000 spectrophotometer (Thermo Fisher). Crystals of GcoA were grown with 2.4 M sodium malonate and 200 mM substrate, dissolved in 40% DMSO where necessary. Crystals were cryocooled directly in liquid N2 without further addition of cryoprotectants. All data were collected at Diamond Light Source (Harwell, UK). For each crystal, 1800 images were taken at 0.1^{o} increments using the default wavelength of 0.9795Å on beamline i04. Data was captured on a Pilatus 6M-F detector. All phases were solved by molecular replacement from the original WT GcoA structure in complex with guaiacol with all non-polypeptide components removed. Data were processed, phased, and models were built and refined using Xia2[455-459] and the Phenix suite.[460-463]

A2.1.7 Strain construction

For genomic manipulations in *P. putida* KT2440, an antibiotic-sacB counter and counterselection method was utilized as described in Blomfield at al.,[464] and modified for *P. putida* KT2440.[465] Diagnostic PCR was utilized to confirm correct integrations or deletions using the 2X myTaq^{TM} system (Bioline). Specific strain construction details are included in Table A2.8, plasmid construction details are included in Table A2.9, and oligo sequences are included in Table A2.10.

To transform cells for episomal expression in *P. putida*, cells were initially grown shaking at 225 rpm, 30 °C, overnight in Luria-Bertani (LB) medium containing 10 g/L

tryptone, 5 g/L yeast extract, and 5 g/L NaCl, and used to inoculate LB the following day to an OD_{600} of ~0.1 and grown until they reached an OD_{600} of 0.6. Cells were then washed three times using 300 mM sucrose and resuspended in a volume 1/50 of the original culture.[466] Plasmid DNA was added, and this mixture was then transferred to 1 mm electroporation cuvettes and electroporated at 1.6 kV, 25 μF, 200 Ω. Following the addition of 950 uL of SOC media (NEB), cells were shaken for an additional 45 minutes at 30°C, 220 rpm, and plated on solid media containing appropriate antibiotics.

A2.1.8 Shake flask experiments and in vivo product analyses

Strains evaluated in shake flask experiments were initially grown overnight in LB media with appropriate antibiotics from glycerol stocks and resuspended the following day in M9 minimal media (6.78 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 0.5 g/L NaCl, 1 g/L NH_4Cl, 2 mM $MgSO_4$, 100 μM $CaCl_2$, and 40 μM $FeSO_4 \cdot 7H_2O$) supplemented with 20 mM glucose (Sigma-Aldrich) and 50 mg/mL kanamycin. Cells were grown until they reached an OD_{600} of 1, at which point syringol (Sigma-Aldrich) was added to a final concentration of 1 mM. Syringol concentrations were quantified using any ^1H NMR, and spectra were collected using a Bruker Avance III HD 400 MHz Spectrometer and analyzed using Bruker TopSpin3.7 software. For identification of unknown compounds, samples were analyzed via LC-MS-MS using the Acquity UPLC system (Waters Inc.). Compounds were separated using a C18 (evo) column (Kinetex) with the mobile phase comprised of a water, acetonitrile, and .1% formate. Flow was directly analyzed by SYNAPT HD-MS using electron spray ionization (ESI) in negative ion mode.

A2.1.9 Bioinformatics Analysis

CYP255A sequences were retrieved by a blastp search against the non-redundant protein sequence database,[467] using GcoA as the query sequence. From the blast results, only sequences with a query cover of at least 66% and sharing a sequence identity of greater than 28% but less than 97% with the query sequence were retained. In total, 482 homologs of GcoA were retrieved. A multiple sequence alignment (MSA) was carried out using MAFFT[144] with Biopython.[468] Conservation analysis was implemented by computing the relative entropy for each site in the MSA.[145] The relative entropy is given by Equation A2.6:

$$\mathrm{R.E} = \sum_{i=1}^{20} (p_i \log \frac{p_i}{p_i^{MSA}}) \qquad \text{(A2.6)}$$

where p_i is the frequency of amino acid i in the given site and p_i^{MSA} is the frequency of amino acid i in the MSA.

A2.1.10 Molecular dynamics (MD) simulations

Unrestrained MD simulations. Unrestrained MD simulations were performed on both WT GcoA and F169A mutant with either guaiacol or syringol bound at the active site. The previous constructed model of WT GcoA with guaiacol bound at the active site was used as the initial conformation to generate these structures.[70] The heme group was kept in a hexacoordinate state (Compound I). Similar to previous work, five histidine residues, including His131, His221, His224, His255, and His343 were doubly protonated (+1 charge), while the remaining histidine residues were kept neutral (singly protonated). Na^+ cations were added to each system to achieve charge neutrality, resulting in ~74,000 atoms for each system.

The ff14SB Amber force field[469] parameters were used for the enzyme and Generalized Amber Force Field (GAFF) parameters[470, 471] were used for the various substrates, as reported previously.[70] Force field parameters for the heme group were taken from ref [472]. Particle mesh Ewald (PME)[473] was used for long-range interactions and the cutoff distance for nonbonded interactions was 9 Å. All the simulations were conducted using NAMD program[474] with a time-step of 2.0 fs. SHAKE algorithm[475] was used to keep bonds to hydrogen atoms fixed. Langevin thermostat with a collision frequency of 1.0 ps^{-1} was used to keep the temperature at 300 K. Each system was relaxed in NPT ensemble for 500 ps first, followed by 80 ns of production was performed in the NVT ensemble. For the initial NPT dynamics, the pressure was held at 1 atm using a Nosé− Hoover barostat coupled to a Langevin piston, [476, 477] with a damping time of 100 fs and a period of 200 fs.

As in our previous work,[70] we employed the following reaction coordinate to describe the opening/closing motion of GcoA: $\xi = RMSD_{open} - RMSD_{closed}$, where $RMSD_{open}$ is the RMSD relative to the most open structure of WT GcoA/apo obtained via microsecond MD trajectory, and $RMSD_{closed}$ is relative to the closed crystal structure. The Cα atoms of residues 5:35 and 154:210 were chosen to calculate the RMSD.

Replica exchange thermodynamic integration (RETI) simulations. In order to assess relative free energies of binding in the active site of GcoA, RETI simulations were performed,[478, 479] using NAMD 2.12.[480] Simulations were performed in the NPT ensemble with simulation parameters identical to those of the MD simulations described above, unless noted otherwise. Initial configurations for RETI were produced from the crystal structure appropriate for each system along with modified AMBER prmtop files to accommodate the dual-topology paradigm employed by NAMD for alchemical

simulations. Each TI replica was simulated for between 150 -250 ns until convergence was achieved. The temperature was maintained at 308 K through Langevin dynamics using a collision frequency of 1.0 ps^{-1}. The cutoff distance for nonbonded interactions was 12.0 Å.

Two separate types of alchemical transitions were performed, with transformations targeting either the substrate or the enzyme. In each case, the transition parameter λ describes the progress of alchemical transformation. In the first kind, a syringol molecule at λ = 0 undergoes an alchemical transition to a guaiacol molecule at λ = 1. This simulation was performed with the substrate molecule in three different contexts: in a periodic water box, in the active site of solvated WT GcoA, and in the active site of solvated GcoA F169A mutant. In the second group of transformations, residue 169 of the GcoA is transformed from Phe at λ = 0 to Ala at λ = 1. This transformation was also performed in three separate contexts: with guaiacol at the active site, with syringol at the active site, and with empty active site (apo). Although these alchemical transitions were made between the entire substrate molecule (in the first set) or protein residue (in the second set), in both cases many atoms are shared in common, so zero-length 10 kcal/mol/$Å^2$ "bonds" were applied between equivalent heavy atoms, "pinning" them together. These additional restraints eliminate sampling of unphysical conformations that can slow the convergence of the calculations.[481-483]

The RETI simulations were performed with 20 windows at λ ∈ [0.00, 0.03, 0.06, 0.10, 0.15, 0.22, 0.29, 0.36, 0.43, 0.5, 0.56, 0.62, 0.68, 0.74, 0.80, 0.86, 0.90, 0.94, 0.97, 1.00}. Smaller increments in λ were used near the end points to avoid "end-point catastrophes".[484] The van der Waals contributions of both disappearing and appearing residues were simultaneously varied with the reaction coordinate λ. Electrostatic

contributions of disappearing residues were turned off over the first half of the reaction coordinate, while those of the appearing residues were turned on over the second half of the reaction coordinate, such that both appearing and disappearing atoms were uncharged at $\lambda = 0.5$. Exchanges between adjacent replicas were attempted every 2 ps between alternating replica pairs, yielding an overall exchange attempt rate of once per 4 ps, consistent with literature guidelines.[485] The combination of these parameters yielded an acceptance ratio for exchanges of approximately 50%. In each case, the final 100 ns were used to compute each ΔG reported in **Figure A2.11**. Third order spline interpolation was used to integrate the average $dU/d\lambda$ obtained from simulations at each window over the final 100 ns. Error estimates were obtained using the methodology detailed in Steinbrecher et al.[486] To avoid underestimating noise error due to autocorrelations of the $dU/d\lambda$ timeseries, means and standard errors were computed by sampling all of the $dU/d\lambda$ data at a rate higher than the output rate. For each simulation, this rate was set by finding the longest triple e-folding time of the $dU/d\lambda$ autocorrelation, as determined by an exponential fit. Typical values of this time were 3-4 ps.

A2.1.11 DFT calculations

Density Functional Theory (DFT) calculations were performed using Gaussian 09.[487] A truncated model containing the porphyrin pyrrole core, Fe center and a methanethiol to mimic cysteine as Fe-axial ligand was used. Geometry optimizations and frequency calculations were performed using unrestricted B3LYP (UB3LYP)[488-490] with the LANL2DZ basis set for iron and 6-31G(d) on all other atoms. Transition states had one negative force constant corresponding to the desired transformation. Enthalpies and

entropies were calculated for 1 atm and 298.15 K. A correction to the harmonic oscillator approximation, as discussed by Truhlar and co-workers, was also applied to the entropy calculations by raising all frequencies below 100 cm^{-1} to 100 cm^{-1}.[491, 492] Single point energy calculations were performed using the dispersion-corrected functional (U)B3LYP-D3(BJ),[493, 494] with the LANL2DZ basis set on iron and 6-311+G(d,p) on all other atoms, within the CPCM polarizable conductor model (diethyl ether, $\varepsilon = 4$) to have an estimation of the dielectric permittivity in the enzyme active site.[495, 496] The use of a dielectric constant $\varepsilon=4$ has been proved to be a good and general model to account for electronic polarization and small backbone fluctuations in enzyme active sites.[497, 498] All stationary points were verified as minima or first-order saddle points by a vibrational frequency analysis. Computed structures are illustrated with CYLView.[499] DFT optimized geometries and DFT Cartesian coordinates of optimized structures are present at the end of this document.

A2.2 Supplementary Figures

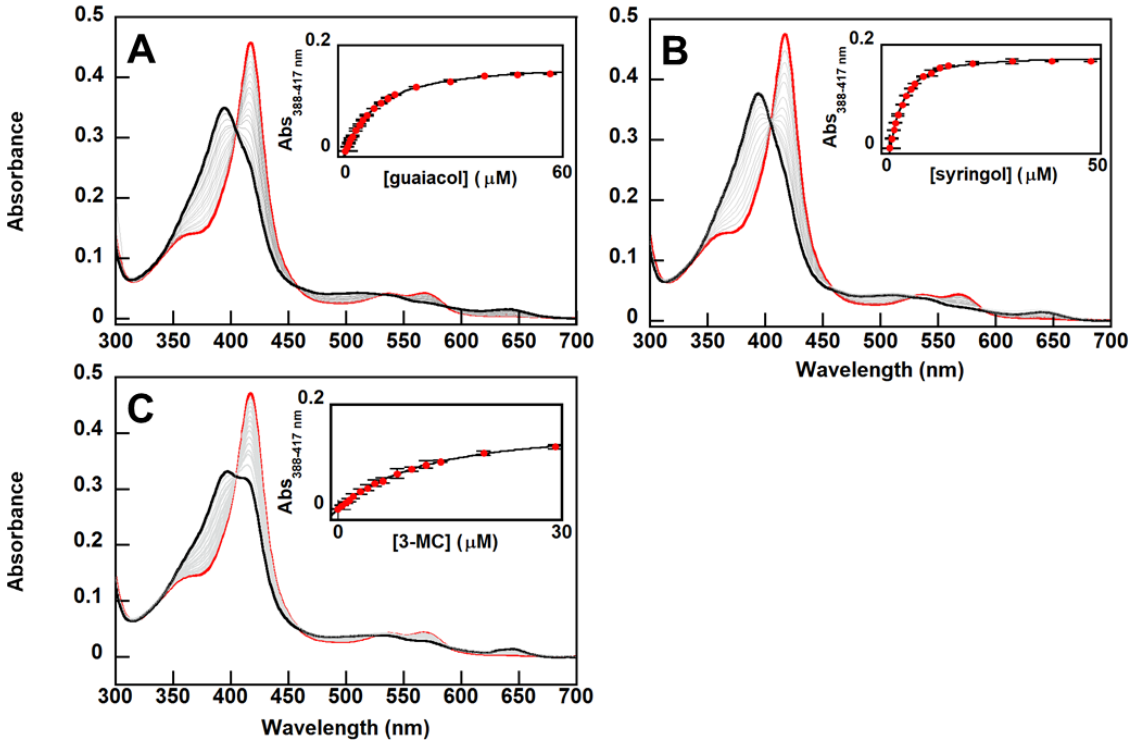


Figure A2.1. Binding of both guaiacol and syringol shows a spin state change in GcoA-F169A and comparable K_Ds**.** Guaiacol (A), syringol (B), and 3MC (C) caused the Soret peak (λ = 417 nm) of GcoA-F169A to shift to 387 nm, indicating the conversion of a low-spin (red trace) to a high-spin (black trace) species. 0-60 μM substrate was titrated into a solution containing 3 μM GcoA-F169A and air-saturated buffer (25 mM HEPES, 50 mM NaCl, pH 7.5, 25 °C) and the spectra monitored until there was no more change, indicating saturation. The solution reached equilibrium prior to each substrate addition. $Abs_{388\text{-}417\ nm}$ was plotted against [substrate] and fit to the quadratic equation for weakly binding ligands (see *SI Appendix,* Methods) to obtain values for K_D: guaiacol = 7.1 ± 0.1 μM, syringol = 1.7 ± 0.07 μM, and 3MC = 9.5 ± 0.02 μM. Error bars represent ±1 standard deviation of three or more runs.

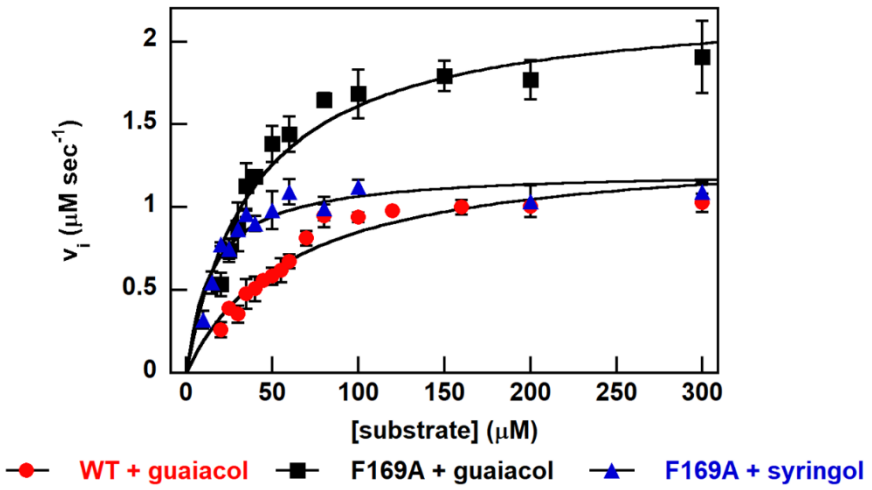


Figure A2.2. GcoA-F169A demethylation of both guaiacol and syringol occurs as or more efficiently than with WT GcoA. Initial rate of NADH consumption is plotted with either GcoA-F169A or WT GcoA as catalyst (0.2 μM) and either guaiacol or syringol as the substrate (300 μM NADH, 100 μg catalase, 210 μM O_2, 25 mM HEPES, 50 mM NaCl, pH 7.5, 25 °C). The data were fit to the Michaelis Menton equation. Error bars represent ±1 standard deviation of three or more runs. WT GcoA is unable to demethylate syringol; hence, only guaiacol data are shown.

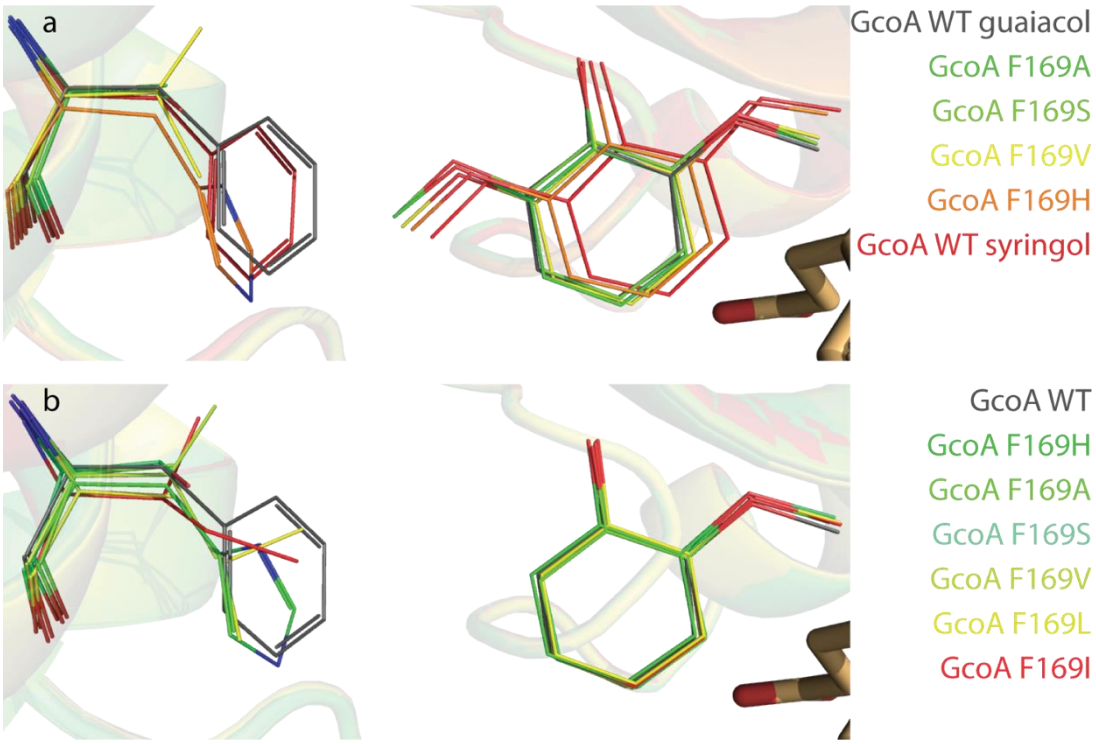


Figure A2.3. Crystallographic analysis for optimal orientation of syringol and guaiacol in multiple complex structures. Superposition of ten ligand-bound mutant structures with the previously characterized WT provided a basis for the determination of the minimum amino acid change required at position 169 for the optimum orientation of syringol, while maintaining a stable environment for guaiacol. The WT position of residue GcoA-F169 is shown in gray in both figures for reference. (A) The syringol molecules in each of the five structures shown are colored on a scale from the unproductive position of the WT in red, improved (orange and yellow), through to the optimum orientation (green) as judged by specific activity (Table A2.4). (B) A superimposition of a total of six alternative mutant structures (GcoA-F169H, A, S, V, L, or I) with the WT indicated that there is no significant change from optimum guaiacol orientation induced by modification at this position.

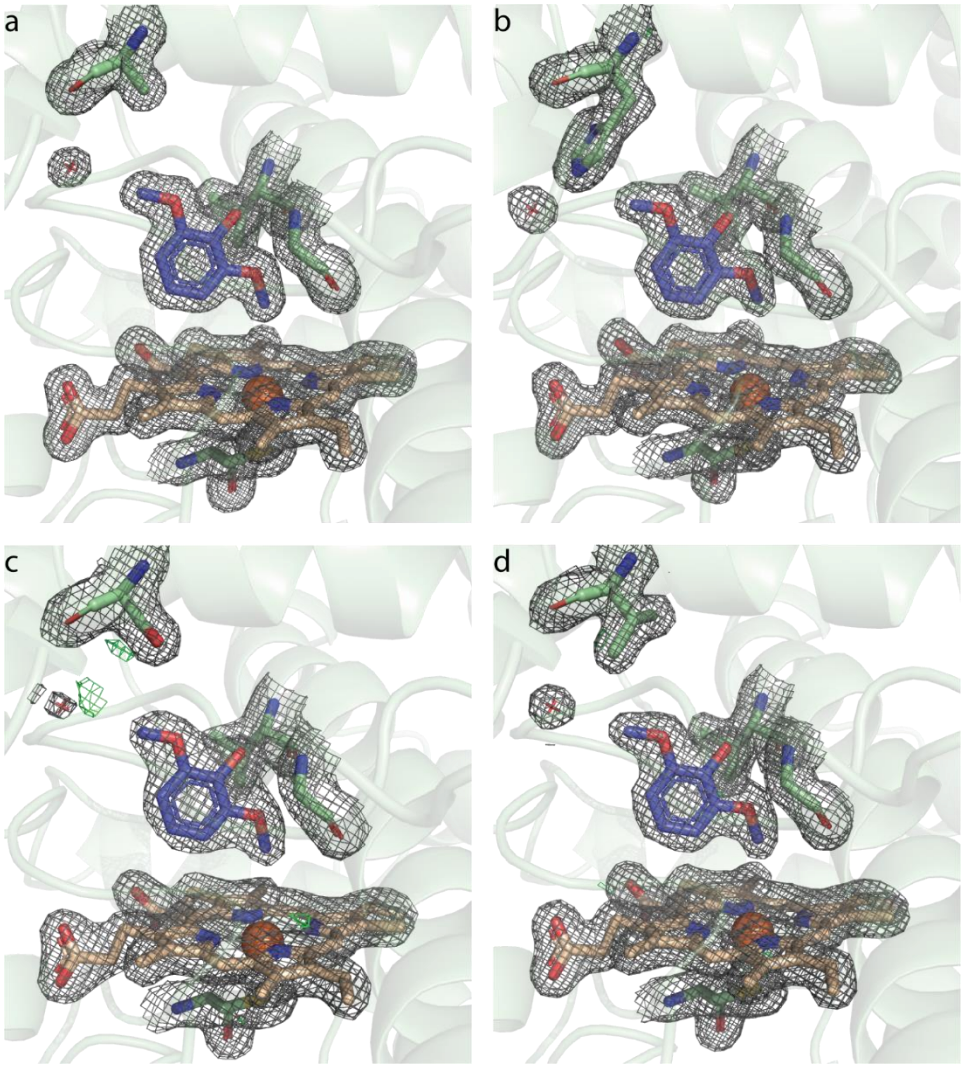


Figure A2.4. Electron density of the active site of GcoA-F169 mutants bound to syringol. Panels (A) – (D) show the structures of syringol bound to GcoA-F169A, H, S and V, respectively.

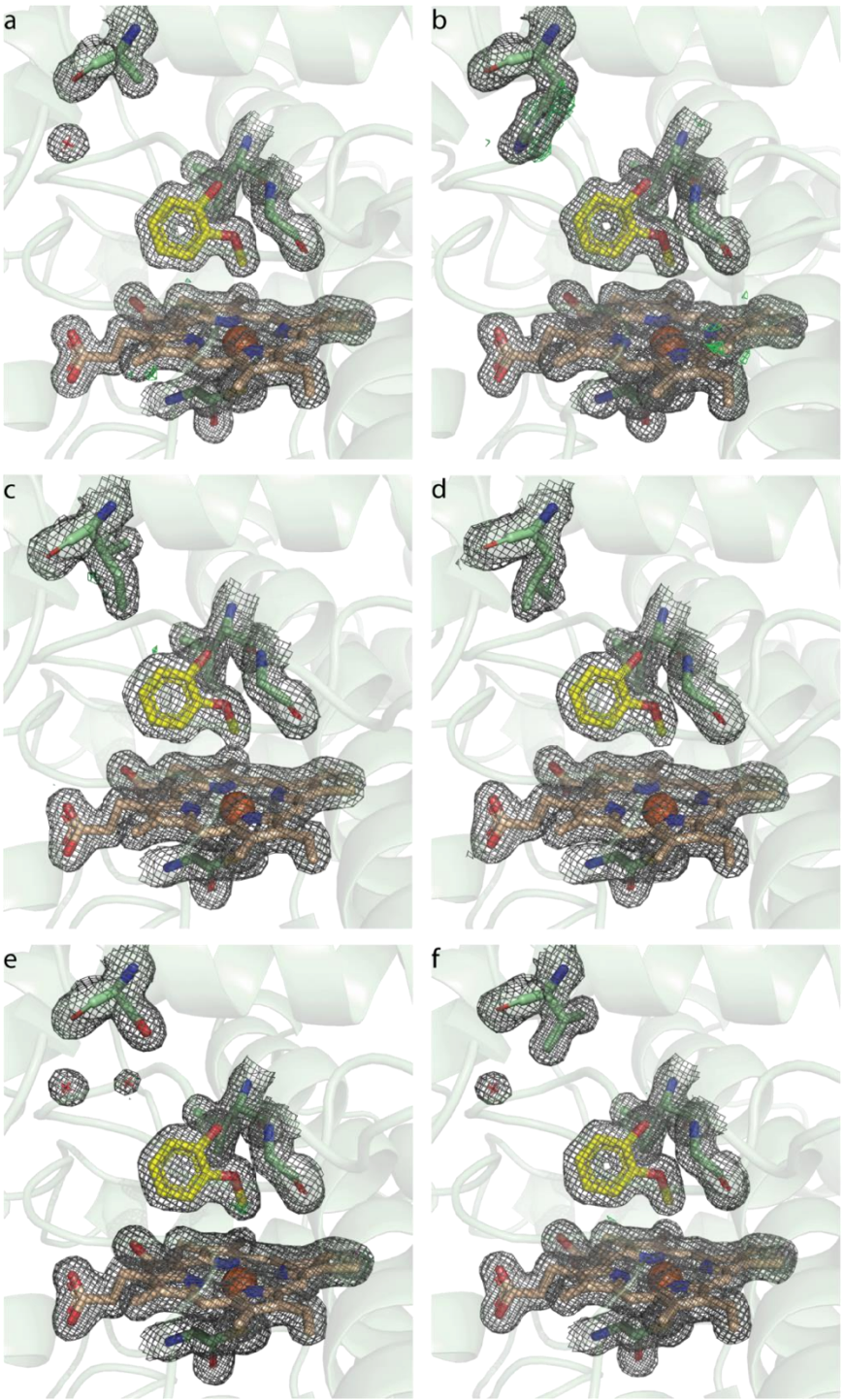


Figure A2.5. Electron density of the active site of GcoA-F169 mutants bound to guaiacol. Panels (A) – (F) show the structures of guaiacol bound to GcoA-F169A, H, I, L, S and V, respectively.

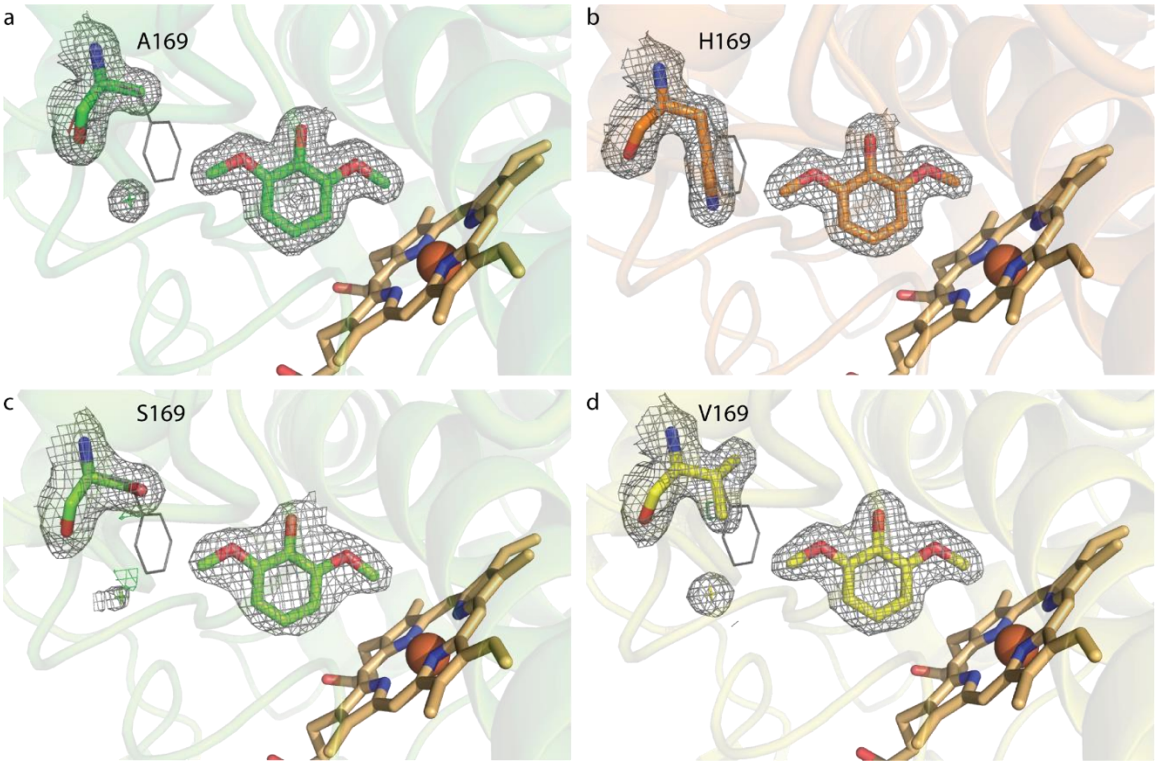
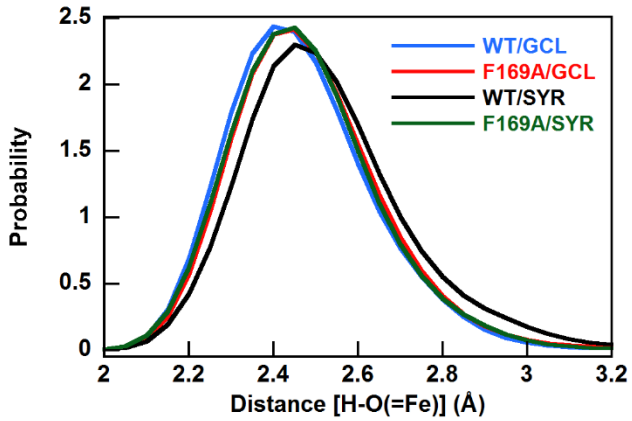


Figure A2.6. Positioning of waters in the active sites of GcoA mutants with syringol. The active sites of three of the syringol bound mutants (GcoA-F169A, S, and V – panels a, c, and d respectively) contain an extra ordered water molecule in the active site. GcoA-F169H does not (panel b). GcoA-F169A bound to syringol additionally contains another water molecule, but it has weak electron density, indicating low occupancy and/or a degree of disorder (not shown). We hypothesize that these ordered water molecules occupy space in the active site that helps to maintain the substrate in a productive binding pose.



GcoA variant	guaiacol	syringol
WT	2.44 ± 0.003 Å	2.50 ± 0.003 Å
F169A	2.46 ± 0.003 Å	2.45 ± 0.003 Å

Figure A2.7. MD simulations indicate that WT GcoA bound with syringol lengthens the key distance for the rate-limiting step for demethylation, namely the distance between the methyl hydrogen and the oxygen atom bound by Compound I heme. The GcoA-F169A mutation brings this distance back to that seen with guaiacol, promoting demethylation. It should be noted that this shift as displayed in MD simulations is significantly subtler than the substrate shift seen in crystal structures (Figure 5.1). Error bars quoted in the table represent the standard error of the mean.

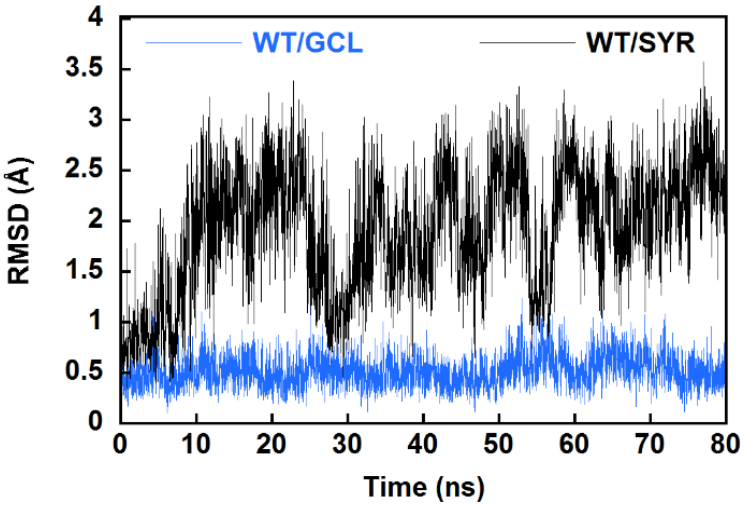
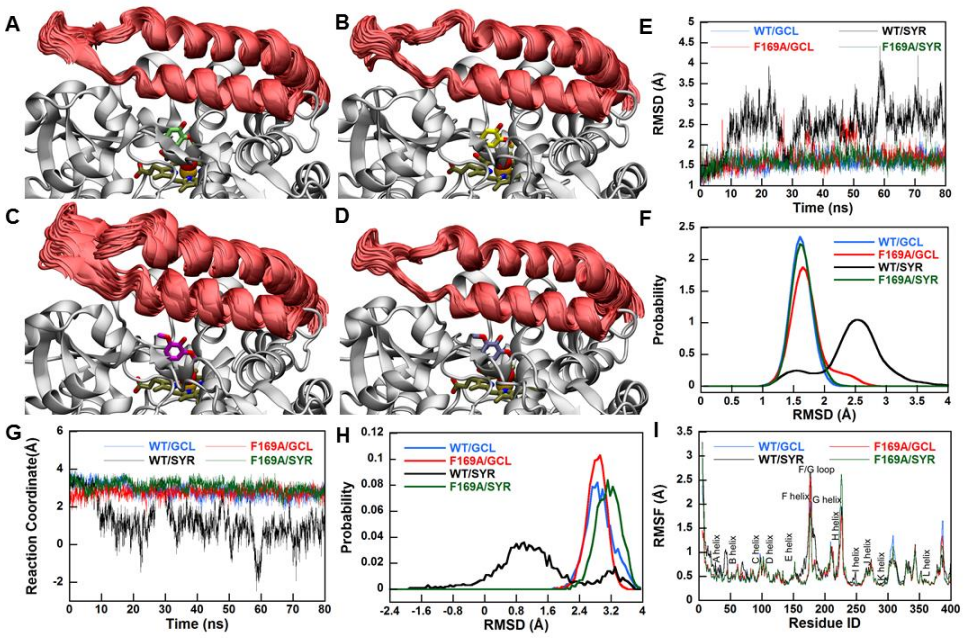


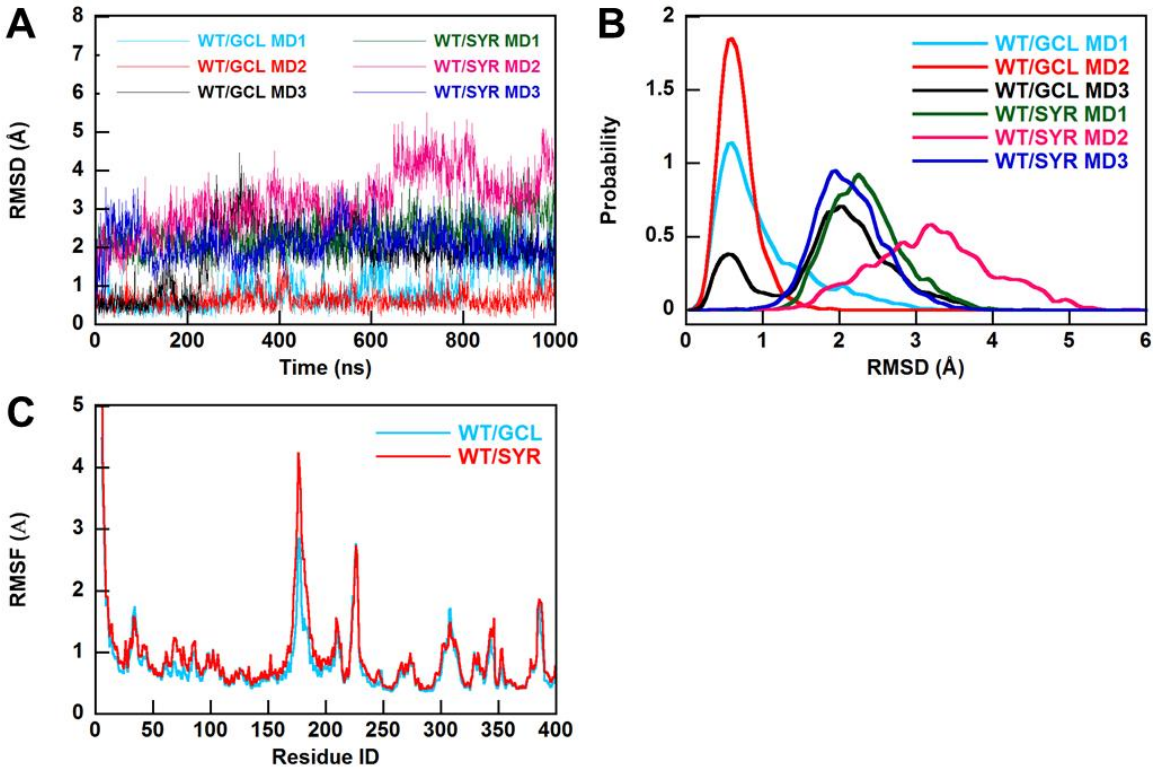
Figure A2.8. MD simulations indicate that introducing syringol into the active site of WT GcoA (rather than guaiacol) results in the displacement of GcoA-F169. Shown is the RMSD of the six ring carbons of GcoA-F169 from their crystal structure positions over the course of 80 ns MD simulation.



	WT/guaiacol	F169A/guaiacol	WT/syringol	F169A/syringol
F169 / A169 (C_α)	0.33	0.37	0.67	0.36
F169 (sidechain)	0.41	n/a	0.98	n/a
F75, F169 / A169, F395 (C_α)	0.39	0.37	0.53	0.37
Ligand	0.44	0.43	0.57	0.47
Heme	0.39	0.43	0.45	0.41
Loop/helix F/G (154:210) (C_α)	0.62	0.66	0.89	0.62

Figure A2.9. MD simulations indicate that the substrate access loop is displaced and more flexible when syringol is bound in the WT enzyme. A) WT GcoA with guaiacol bound, B) GcoA-F169A GcoA with guaiacol bound, C) WT GcoA with syringol, and D) GcoA-F169A GcoA with syringol bound. Shown in "sticks" are the substrate and heme; heme Fe and Fe-bound oxygen are shown as spheres. The substrate access loop (residues 154:210) is shown in "cartoon" representation every 2 ns over the course of the 80 ns MD simulation. E) RMSD of all heavy atoms of the substrate access loop, as compared to the crystal structure position and F) probability distribution of the same. G) Time trace of the reaction

coordinate for opening and closing of the substrate access loop (negative values indicate a more open architecture) and H) probability distribution of the same. I) RMSF (root mean square fluctuation) of GcoA (alpha carbons only) over the course of each 80 ns MS simulation. The greatest area of difference is seen in the substrate access loop region, with the fluctuations in WT GcoA with syringol bound are seen to be the greatest. The table highlights the RMSF (Å) of selected residues and regions of GcoA, as well as ligands. In this table, the RMSF calculation for "Ligand" and "Heme" is performed on all heavy atoms of those molecules (i.e. no hydrogen atoms). For the entries related to the GcoA enzyme, the RMSF calculation is performed on the alpha carbons only, with one exception: the line for "F169 (sidechain)", the RMSF calculation is for all heavy sidechain atoms.



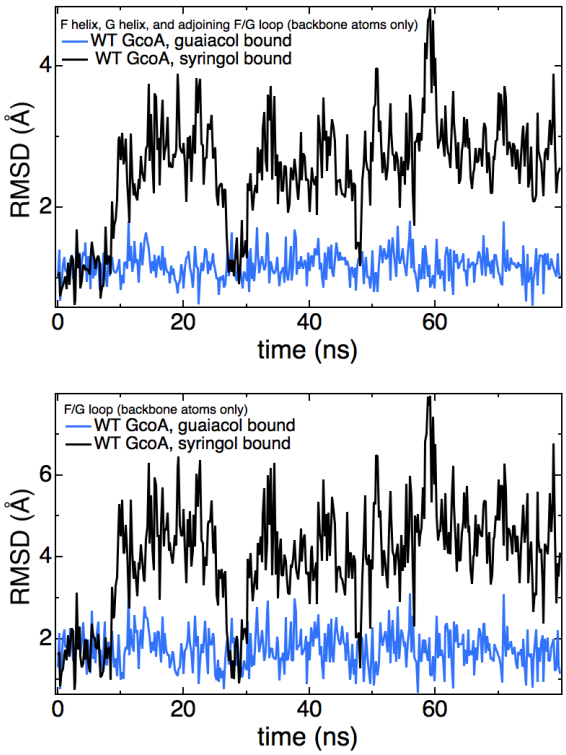
	WT/guaiacol			WT/syringol		
	MD1	MD2	MD3	MD1	MD2	MD3
F169 (sidechain)	0.84	0.54	1.18	0.91	1.83	0.78
F75, F169, F395 (C_α)	0.54	0.44	0.71	0.62	1.12	0.58
Ligand	0.56	0.53	1.60	0.64	1.79	0.58
Heme	0.53	0.50	0.56	0.58	0.69	0.53
Loop/helix F/G (154:210) (C_α)	0.99	0.77	1.21	1.37	1.66	1.05

Figure A2.10. Microsecond MD simulations indicate the increased flexibility and displacement of F169 in WT GcoA when syringol is bound rather than guaiacol. These analyses are performed on three simulations (each) of WT GcoA with guaiacol and with syringol. These simulations were performed in our original study of GcoA;[70] however our previous publication focused on the utilization of guaiacol and presented the results with syringol only very briefly (the fifth entry of Figure A2.21). Time traces of the reaction

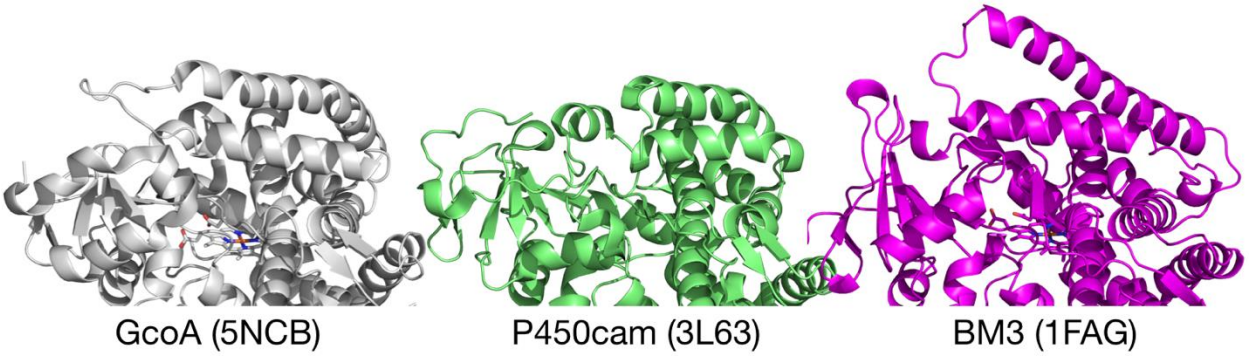
coordinate (RC) for opening and closing as well as the RMSD of the three active site Phe residues were presented in the supplementary section of the prior publication. A) Time traces of the RMSD of residue F169 and B) probability distributions of the same indicate that F169 deviates substantially from its crystal structure position when syringol is bound. The movement of F169 is somewhat correlated with the motion of the substrate access loop. The probability distribution shows that each guaiacol simulation spends significant time with F169 very near the crystal structure position (first peak) whereas none of the syringol simulations display this peak. C) RMSF averaged over three simulations each for WT GcoA with either guaiacol or syringol. The significant area of deviation is in the substrate access loop (which includes F169), indicating greater dynamic flexibility when syringol is bound. Table presents RMSF (Å) of selected residues and regions of GcoA, as well as ligands. In this table, the RMSF calculation for "Ligand" and "Heme" is performed on all heavy atoms of those molecules (i.e. no hydrogen atoms). For the entries related to the GcoA enzyme, the RMSF calculation is performed on the alpha carbons only, with one exception: the line for "F169 (sidechain)", the RMSF calculation is for all heavy sidechain atoms.

P450cam has been captured in both open (e.g. PDB codes 3L61, 3L62) and closed (e.g. 3L63) states.[288] Lee *et al.* note that the RMSD between the closed and open conformation is approximately 4-5 Å for the F helix and G helix and peaked around 8.5 Å at the F/G loop (Figure 2 in that publication). In our 80 ns MD simulations, the RMSD of the combined F and G helices, as well as the adjoining loop fluctuate around an RMSD value of 3 Å, whereas the adjoining loop alone fluctuates around 5 Å, as shown in the

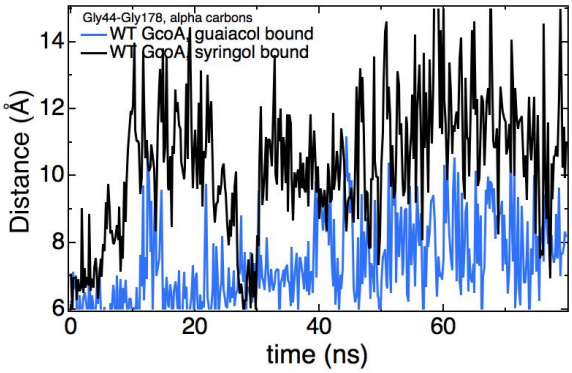
graphs here. Thus, the degree of opening by this metric is on the same order, but lower by about 50% from that observed in P450cam crystal structures.



In addition to P450cam, we also note P450 BM3, which also displays crystallographic evidence of open (e.g. PDB codes 1BVY[289]) and closed (e.g. 1JPZ[290] and 1FAG[500]) structures. Dubey *et al.* presented MD simulations of BM3 in which the opening/closing of access loops was shown to be regulated by substrate binding.[501] Below, we show an aligned structural comparison of GcoA, P450cam, and BM3 in their closed states.



Dubey *et al*. utilize a single distance from alpha carbons to describe the opening motion of BM3 in their simulations (this distance involves a loop that is opposite of the F/G loop). Follmer *et al*. also present residue-residue distance as a metric for channel opening in P450cam.[502] In order to compare our degree of openness for GcoA with those seen in the crystal structures of open/closed forms of P450cam and BM3, we have chosen a metric for each case that stretches from the F/G loop to a loop that is across the substrate access channel. For BM3, this metric is the distance from the alpha carbon of Pro45 to that of Pro196, which ranges from 19.5 Å in the closed (1JPZ) to 26.9 Å in the open state (1BVY), giving a difference of 7.4 Å. For P450cam, an analogous distance is from the alpha carbon of Asn59 to that of Ser190, which ranges from 17.3 Å in the closed structure (3L63) to 25.5 Å in the open structure (3L61), giving a difference of 8.2 Å. For GcoA, a similar metric is the distance between alpha carbons of Gly44 and Gly178, which is 6.35 Å in the closed state (5NCB). In our simulations of WT GcoA, the difference between the simulation with syringol bound and that with guaiacol bound is slightly more than 3 Å. As shown in the graph below, the average distance over the final 40 ns of the simulation is 8.0 Å (guaiacol) and 11.1 Å (syringol).

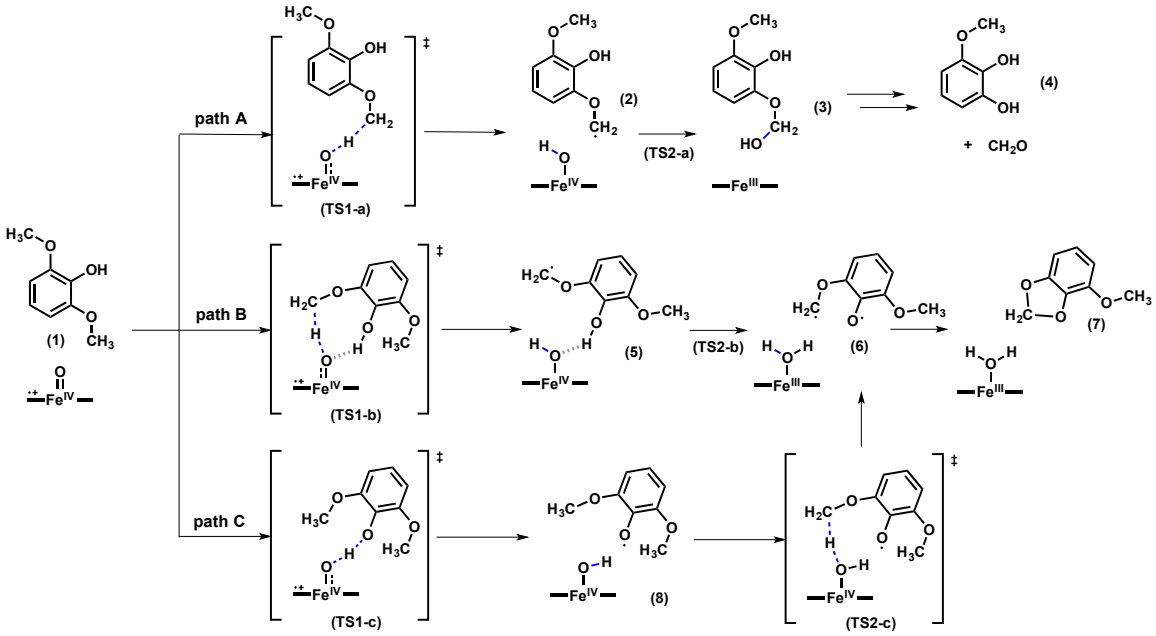


We note that, to date, efforts to crystallize GcoA in the apo state have proven unsuccessful. As a result, all existing crystal structures demonstrate GcoA in the closed state. Thus, the range of motion of the

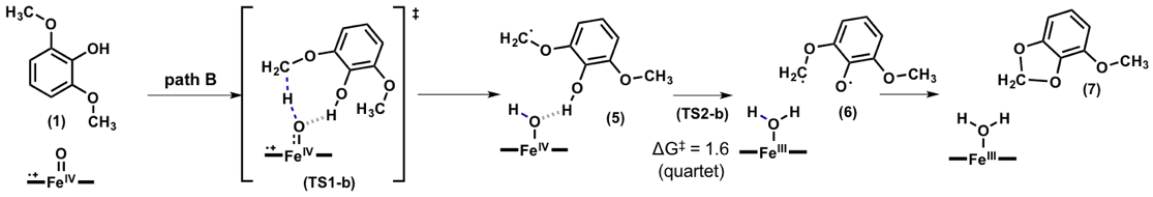
substrate access lid for GcoA is currently unverified experimentally. The range of motion that we observe in our 80 ns simulations is not necessarily the full range of motion in the catalytic cycle of GcoA (and very likely is not). Additionally, it is possible that the primary effect of the increased flexibility and propensity for the substrate access lid to open when syringol is bound could be upon substrate binding/egress or on the catalytic demethylation step.

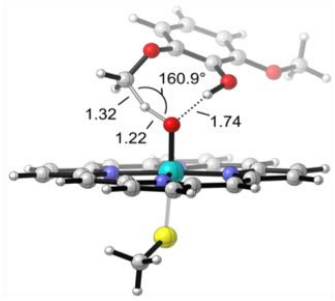
In our 80 ns MD simulations, we find that the substrate access loop displays opening (whether by an RMSD metric or cross-channel distance) within the first 10 ns and remains open for the remaining 70 ns, with only one very brief excursion to the closed state around 30 ns. Focusing on the F/G loop backbone atoms and taking RMSD=3.0 Å as the cutoff between the open and closed states, the simulation with syringol is in the open state for 81% of the simulation, whereas the guaiacol-bound simulation is open for less than 1% of the simulation time. (We note again, that this definition of "open" does not signify that this is the most open state of GcoA).

A)



B)

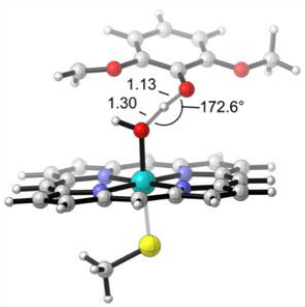




TS1-b²

ΔG‡ = 15.7

Rate limiting TS C-H abstraction



TS2-b⁴

ΔG‡ = 1.6

C)

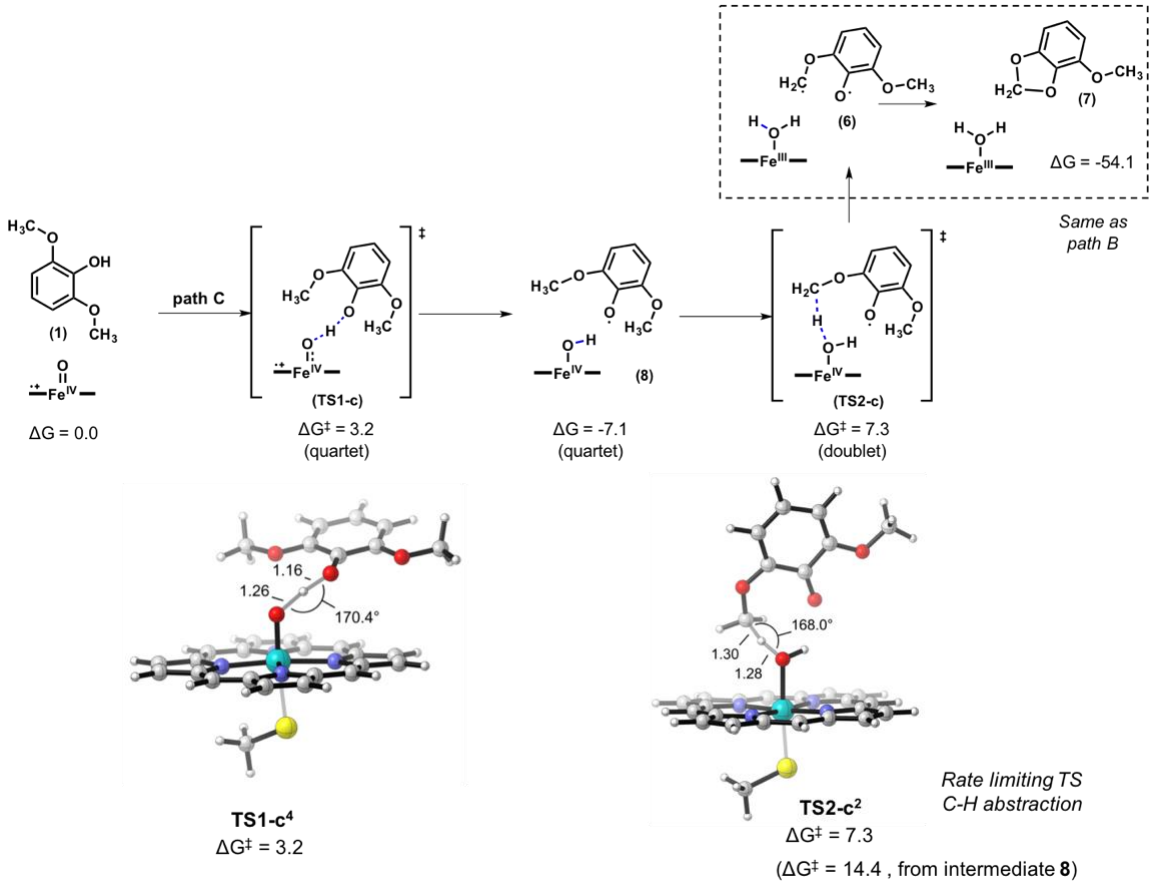


Figure A2.11. DFT calculations for possible reaction pathways for syringol degradation catalyzed by Compound I (CpI). A) Target pathway for syringol demethylation to 3MC catalyzed by P450 CpI. Our calculations indicate syringol demethylation follows a similar pathway to that previously described for guaiacol.[70] Free energy barriers for the rate-limiting initial hydrogen atom transfer (HAT) from the methoxy group to the heme-bound oxygen atom are very similar for guaiacol and syringol ($\Delta G^{\ddagger}$ = 18.6 and 18.9 kcal·mol^{-1}, respectively), and also are the optimized transition state (TS) geometries. Consequently, the demethylation reaction likely proceeds similarly for both substrates when they can bind productively with the CpI reactive species in the enzyme active site. Unproductive pathways b (B) and c (C) lead to undesired products (and possibly uncoupling) and are

intrinsically lower in energy than pathway **a**. This is indicating that enzymatic environment that forces a specific substrate binding pose, as demonstrated by X-ray structures and MD simulations, is preventing pathways **b** and **c** to take place and promotes desired pathway **a**. Gibbs energies, obtained at

B3LYP-D3BJ/6-311+G(d,p)+Fe(LanL2DZ)(PCM=Diethylether)//B3LYP/6-31G(d)+Fe(LanL2DZ),

are given in kcal/mol. Distances and angles in DFT optimized key TSs are given in Å and degrees, respectively.

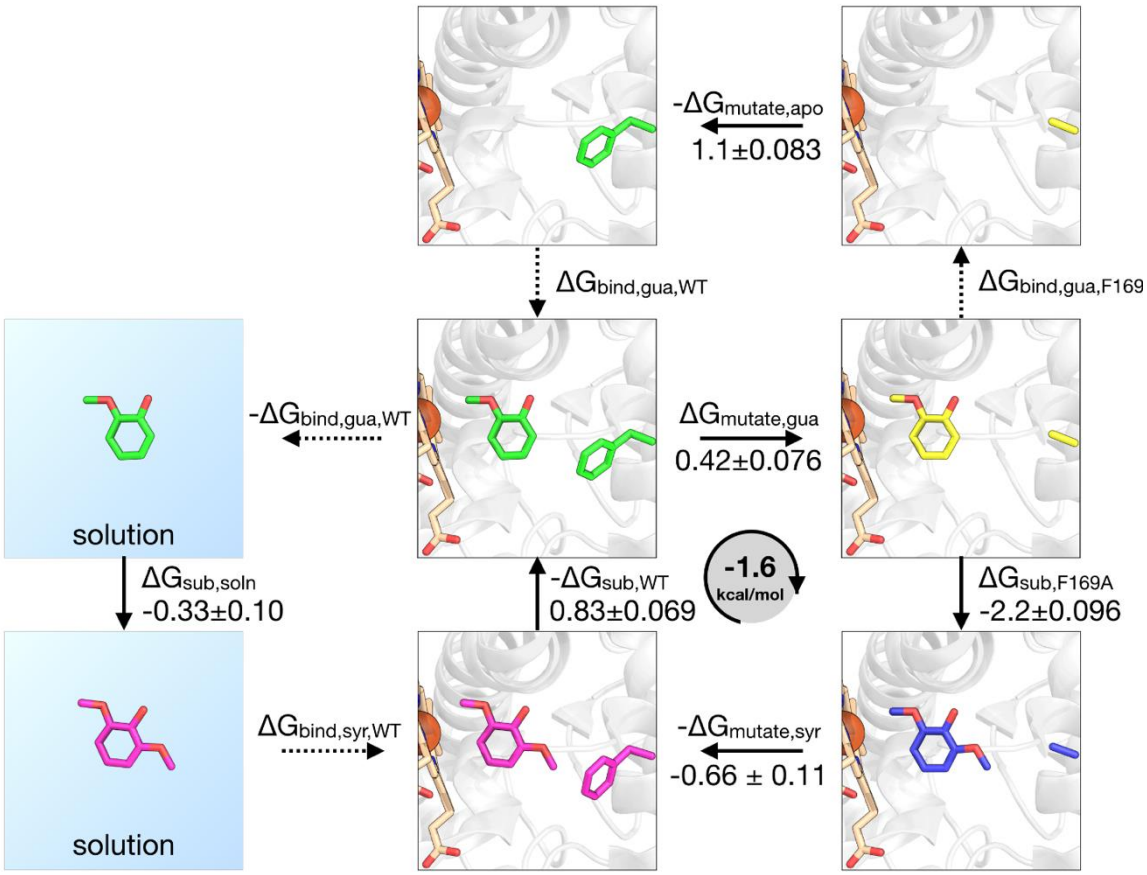


Figure A2.12. Replica exchange thermodynamic integration (RETI) simulations indicate the change in free energy for different alchemical transformations. Each solid arrow indicates a RETI calculation that was performed and has the associated free energy change for the process in the direction of the arrow. Dashed arrows represent transformations that were not performed. A full four-step thermodynamic cycle was completed, represented by the four boxes in the lower right. The closure in performing these four steps theoretically is equal to zero. We calculate a closure of -1.6 kcal/mol, which is on the order of the expected cumulative error one would expect given the errors in the individual steps (approximately 0.2 kcal/mol).

The most surprising of these thermodynamic results is the relatively low free energy difference between syringol and guaiacol bound at the active site of WT GcoA. This can explained by the fact that these simulations indicate that binding syringol in the WT can produce binding complexes in which the substrate is flipped "upside down" from the configuration seen in crystal structures (*see* Figure. A2.12). This may help explain the experimental result that while WT demethylation activity is greatly reduced when exchanging syringol for guaiacol, the binding preference is less dramatic. Substrate flipping could reduce the crowding penalty in WT GcoA. The analogous set of simulations in the GcoA-F169A mutant (as well as mutation transformations with either syringol or guaiacol bound) showed comparatively few flips (for substrate transformation in WT, the substrate was flipped ~90% of the time, compared to ~10% in the other three cases). Experimentally, syringol binding affinity is not much changed from the WT to GcoA-F169A (K_D values indicate a difference in binding free energy of 0.3 kcal·mol^{-1}), even as the activity is improved markedly. The RETI results indicate the possibility that this may be because K_D measurements detect all binding at the active site capable of displacing a Fe-coordinating water molecule whereas catalytic turnover will only result from productive binding with catalytically relevant positioning of substrate relative to CpI heme. RETI results present the possibility that this is because the GcoA-F169A mutant is able to keep both substrates oriented correctly whereas WT GcoA is not.

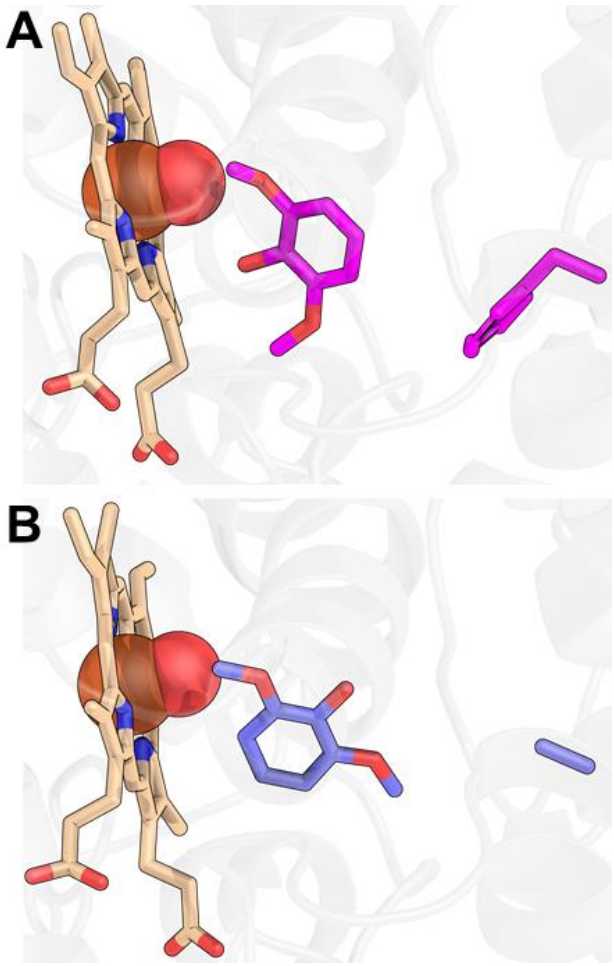


Figure A2.13. Alternate substrate binding configurations seen in RETI simulations. These results indicate a flip of substrate that is unique to the $\Delta G_{sub,WT}$ calculation, which involves an alchemical transformation between guaiacol and syringol in the active site of WT GcoA. Given the differences in K_D, activity, and crystal structures, the expected result for this transformation would be that guaiacol would be significantly favored at the WT active site. It is therefore surprising that this transformation results in a free energy change that slightly *favors* syringol (by 0.83 ± 0.069 kcal/mol); this is explained by the fact that the substrate adopts a unique binding pose not seen in crystal structures or any other MD simulations. A) Exemplified by syringol, in all 20 windows (in which syringol is gradually "turned off" and guaiacol is gradually "turned on") of the $\Delta G_{sub,WT}$ transformation, this "upside down"

configuration dominates, spending only about 10-15% of the simulation time in a configuration similar to that seen in crystal structures. By contrast, in the other transformations ($\Delta G_{mutate,gua}$, $\Delta G_{sub,F169A}$, and $\Delta G_{mutate,syr}$) the "upright" substrate configuration is displayed in upwards of 90% of the simulation time. This is represented in panel B by a representative frame from the $\Delta G_{sub,F169A}$ simulation. Heme, substrate, and sidechain of residue 169 (Phe in panel A and Ala in panel B) are shown in sticks; heme Fe and Fe-bound oxygen are shown as spheres.

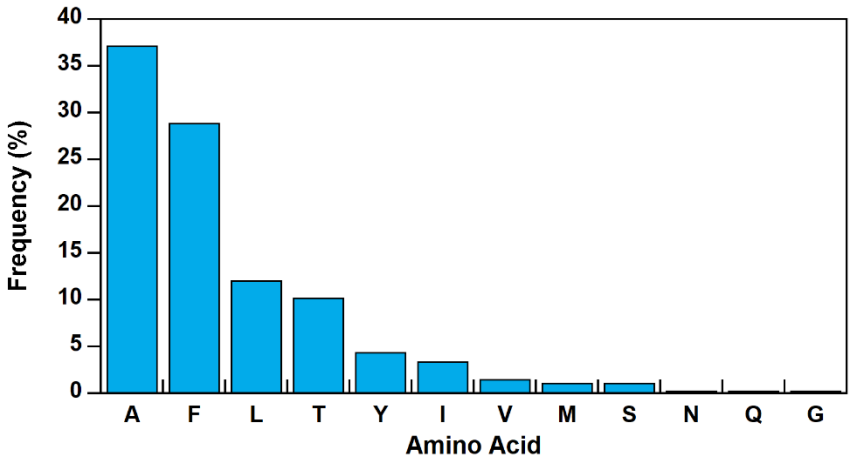


Figure A2.14. Frequency of amino acids occurring at position 169 among GcoA homologs. Ala is more frequent than Phe, and none of the 482 homologs utilizes His at the 169th position.

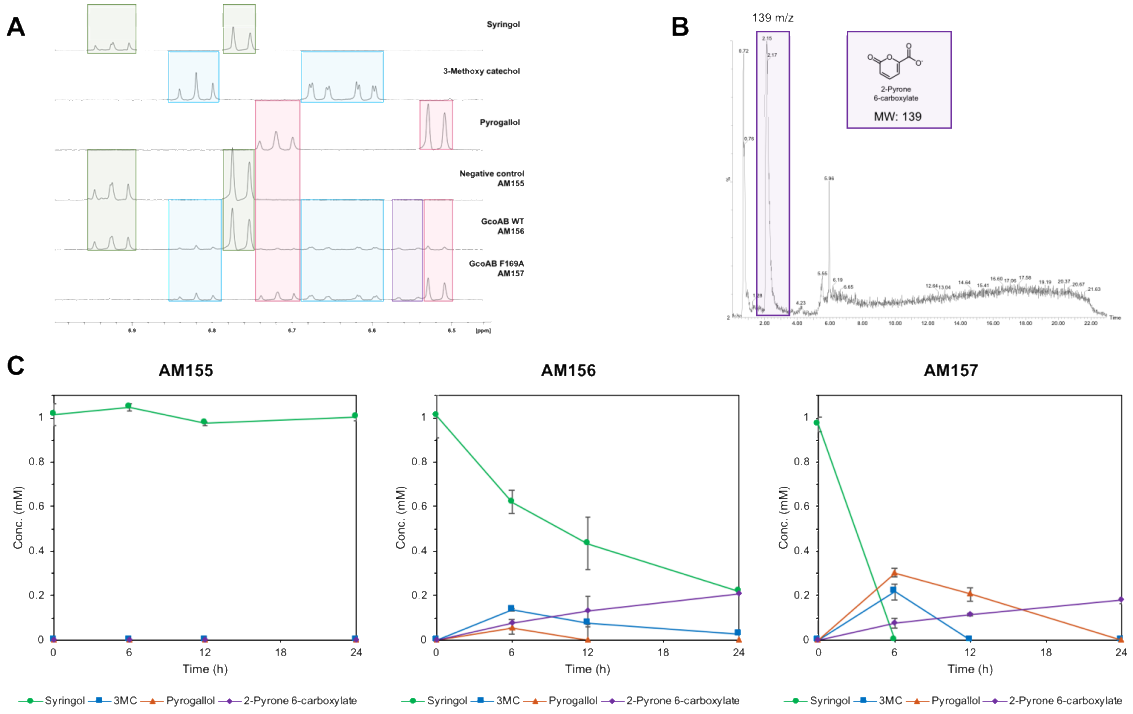


Figure A2.15. *In vivo* syringol consumption. (A) Expanded figure including standards of syringol, 3MC, and pyrogallol. (B) Extracted ion chromatograph of 139 m/z verifying presence of 2-pyrone 6-carboxylate in AM157. (C) Time course analysis of syringol depletion and product formation. Absolute analyte concentrations were measured via integration of ^{1}H NMR peaks, by comparison to an integrated standard peak (TMSP). Plasmid-based expression of $GcoAB_{F169A}$ (AM157) allows for complete turnover of syringol after 6 hours, while syringol still remains after 24 hours in cells expressing WT GcoAB (AM156). Data points represent averages of three replicates, and error bars indicate the standard deviations of the measurements.

A2.3 Supplementary Tables

Table A2.1. Rate of substrate disappearance or product appearance (specific activity) for F169A.

Substrate	NADH disappearance ($\mu M\ s^{-1}\ \mu mol\ F169A^{-1}$)	aromatic substrate disappearance ($\mu M\ s^{-1}\ \mu mol\ F169A^{-1}$)	K_D (μM)	
			WT	F169A
guaiacol	5.8 ± 0.8	5.6 ± 0.3	0.0060 ± 0.002	7.1 ± 0.1
syringol	5.1 ± 0.8	4.1 ± 0.3	2.8 ± 0.4	1.7 ± 0.07
3MC	5.6 ± 0.8	2.6 ± 0.3	3.7 ± 0.1	9.5 ± 0.02

Table A2.2. Rates of substrate disappearance and coupling efficiencies for reactions catalyzed by the GcoA-F169 variants**.**

[a]NADH consumption was monitored continuously over time via UV/vis and quantifying with ε_{340} = 6.22 $mM^{-1}\ cm^{-1}$ at 25°C, 25 mM HEPES, 50 mM NaCl, pH 7.5 and saturating (200 μM) concentrations of all substrates (syringol, guaiacol, and 3MC).

[b]Calculated as the molar ratio of formaldehyde produced per NADH consumed in a fixed-time assay. Assay conditions: 0.2 μM GcoA variant, 0.2 μM GcoB, 200 μM NADH, 100 μg/mL catalase, 200 μM aromatic substrate in 25 mM HEPES, 50 mM NaCl, pH 7.5, 25°C, 210 μM O_2.

	syringol		guaiacol		3MC	
GcoA variant	Specific activity ($\mu mol\ sec^{-1}\ \mu mol^{-1}$ GcoA)[a]	Coupling efficiency (%)[b]	Specific activity ($\mu mol\ sec^{-1}\ \mu mol^{-1}$ GcoA)[a]	Coupling efficiency (%)[b]	Specific activity ($\mu mol\ sec^{-1}\ \mu mol^{-1}$ GcoA)[a]	Coupling efficiency (%)[b]
WT	n/a	7.1 ± 0.8	5.0 ± 0.1	110 ± 10	3.2 ± 0.2	78 ± 3
F169A	5.1 ± 0.8	104 ± 6	5.8 ± 0.8	103 ± 7	5.6 ± 0.8	64 ± 10
F169S	5.1 ± 0.9	85 ± 5	5.7 ± 0.8	67 ± 8	6.0 ± 0.8	67 ± 8
F169V	4.1 ± 0.8	100 ± 10	5.3 ± 0.2	105 ± 2	5.7 ± 0.3	40 ± 3
F169H	3.9 ± 0.2	56 ± 7	7.9 ± 3	103 ± 10	4.3 ± 0.4	28 ± 20
F169I	0.56 ± 0.2	14 ± 2	1.4 ± 0.2	41 ± 10	1.1 ± 0.7	10 ± 9
F169L	0.54 ± 0.1	7.8 ± 2	4.5 ± 0.3	73 ± 3	0.57 ± 0.2	5.0 ± 4

Table A2.3. Uncoupling reactions with GcoA-F169A and guaiacol, syringol, or 3MC**.**

[a]0.2 µM GcoA-F169A and GcoB were reacted with 100 µg catalase, 100 or 200 µM NADH, and 100 µM substrate in 25 mM HEPES, 50 mM NaCl, pH 7.5, 25 °C, 210 µM O_2. Endpoint analyses were done to quantify formaldehyde produced (see Methods), which was then referenced to NADH consumed.

[b]The same reaction done above was repeated to quantify H_2O_2 produced using the Amplex Red/HRP assay (see *SI* Appendix, Methods). The amount of hydrogen peroxide was then referenced to NADH consumed.

Substrate	% Formaldehyde produced[a]	% H_2O_2 produced[b]
guaiacol + 100 µM NADH	100 ± 10	3.7 ± 0.6
syringol + 100 µM NADH	125 ± 6	7.4 ± 0.03
syringol + 200 µM NADH	91 ± 8	7.9 ± 0.5
3-MC + 100 µM NADH	52 ± 7	17 ± 0.2

Table A2.4. X-ray tables for GcoA syringol-bound structures

PROTEIN	F169A WITH SYRINGOL	F169H WITH SYRINGOL	F169S WITH SYRINGOL	F169V WITH SYRINGOL
PDB CODE	6HQQ	6HQR	6HQS	6HQT
DATA COLLECTION				
SPACE GROUP	P43212	P43212	P43212	P43212
WAVELENGTH (Å)	0.9795	0.9795	0.9795	0.9795
CELL DIMENSION				
A, B, C (Å)	104.07, 104.07, 116.10	102.15, 102.15, 109.98	104.25, 104.25, 114.03	103.90, 103.90, 115.58
A, B, Γ (˚)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
RESOLUTION RANGE (Å)	77.49-1.66 (1.70-1.66)	72.23-1.79 (1.84-1.79)	61.91-2.17 (2.23-2.17)	62.00-1.85 (1.90-1.85)
RMERGE	0.066 (1.244)	0.081 (0.938)	0.066 (0.816)	0.070 (0.813)
RMEAS	0.071 (1.351)	0.087 (1.018)	0.072 (0.888)	0.075 (0.917)
RPIM	0.027 (0.525)	0.034 (0.393)	0.028 (0.347)	0.029 (0.415)
I/ΣI	19.0 (2.0)	15.8 (2.2)	20.9 (3.0)	22.5 (2.3)
COMPLETENESS (%)	100 (100)	100.0 (100.0)	100.0 (100.0)	99.9 (99.9)
MULTIPLICITY	12.9 (12.7)	12.5 (12.7)	12.7 (12.3)	12.7 (8.9)
CC 1/2	1.00 (0.794)	0.999 (0.878)	0.999 (0.865)	1.000 (0.809)
REFINEMENT				
RESOLUTION RANGE (Å)	73.59-1.66 (1.69-1.66)	51.05-1.79 (1.82-1.79)	52.15-2.17 (2.23-2.17)	50.50-1.85 (1.88-1.85)
NO. OF REFLECTIONS	75469	53703	33716	54478
RWORK	0.1676 (0.3303)	0.1692 (0.2999)	0.1608 (0.2507)	0.1563 (0.2428)
RFREE	0.1909 (0.3748)	0.1991 (0.3939)	0.1858 (0.2932)	0.1788 (0.2987)
NO. OF ATOMS	3752	3629	3411	3601
PROTEIN	3151	3157	3122	3154
LIGAND/ION	54	54	54	54
WATER	547	418	235	393
B-FACTORS	21.9	21.97	37.01	27.66
PROTEIN	27.19	24.89	41.15	29.53
LIGAND/ION	18.06	15.15	28.99	21.89
WATER	41.41	37.93	45.92	41.32
RMSD				
BOND LENGTHS (Å)	0.008	0.016	0.008	0.007
BOND ANGLES (˚)	1.012	1.353	0.938	0.9

Table A2.5. X-ray tables for GcoA guaiacol-bound structures

PROTEIN	F169A WITH GUAIACOL	F169H WITH GUAIACOL	F169I WITH GUAIACOL	F169L WITH GUAIACOL	F169S WITH GUAIACOL	F169V WITH GUAIACOL
PDB CODE	6HQK	6HQL	6HQM	6HQN	6HQO	6HQP
DATA COLLECTION						
SPACE GROUP	P43212	P43212	P43212	P43212	P43212	P43212
WAVELENGTH (Å)	0.9795	0.9795	0.9795	0.9795	0.9795	0.9795
CELL DIMENSION						
A, B, C (Å)	103.67, 103.67, 114.88	102.00, 102.00, 109.83	105.20, 105.20, 112.52	104.00, 104.00, 111.58	103.85, 103.85. 115.10	103.80, 103.80, 115,78
A, B, Γ (˚)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
RESOLUTION RANGE (Å)	76.97-1.57 (1.61-1.57)	60.29-1.49 (1.53-1.49)	56.25-1.85 (1.90-1.85)	76.08-1.87 (1.92-1.87)	73.43-1.70 (1.74-1.70)	73.40-1.62 (1.66-1.62)
RMERGE	0.071 (0.964)	0.077 (1.040)	0.064 (1.002)	0.080 (0.991)	0.064 (1.309)	0.069 (0.985)
RMEAS	0.077 (1.068)	0.080 (1.040)	0.069 (1.085)	0.087 (1.076)	0.069 (1.420)	0.075 (1.096)
RPIM	0.029 (0.457)	0.032 (0.449)	0.027 (0.415)	0.034 (0.417)	0.026 (0.547)	0.029 (0.343)
I/ΣI	18.3 (2.2)	14.9 (2.3)	10.4 (2.6)	17.0 (2.4)	24.0 (2.0)	17.4 (2.1)
COMPLETENESS (%)	100 (100)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
MULTIPLICITY	12.5 (10.5)	12.6 (12.2)	12.7 (13.1)	12.7 (12.7)	13.1 (12.9)	12.4 (10.2)
CC 1/2	0.999 (0.802)	0.999 (100.0)	1.000 (0.849)	0.999 (0.808)	1.000 (0.786)	0.999 (0.828)
REFINEMENT						
RESOLUTION RANGE (Å)	76.98-1.57 (1.59-1.57)	60.00-1.49 (1.51-1.49)	52.61-1.85 (1.88-1.85)	73.54-1.87 (1.91-1.87)	73.43-1.70 (1.72-1.70)	73.40-1.62 (1.64-1.62)
NO. OF REFLECTIONS	87452	93998	54123	51138	69625	80413
RWORK	0.1581 (0.2897)	0.1744 (0.3340)	0.1610 (0.3039)	0.1455 (0.2439)	0.1517 (0.2641)	0.1661 (0.3377)
RFREE	0.1812 (0.3208)	0.1984 (0.3800)	0.1799 (0.3934)	0.1697 (0.2627)	0.1811 (0.3150)	0.1874 (0.3602)
NO. OF ATOMS	3744	3655	3578	3650	3684	3697
PROTEIN	3161	3157	3087	3158	3144	3154
LIGAND/ION	52	52	52	52	52	52
WATER	531	446	439	440	488	491
B-FACTORS	16.87	17.15	27.3	33.44	25.32	19.35
PROTEIN	20.56	21.82	28.17	35.61	29.52	23.73
LIGAND/ION	13.52	13.31	18.97	25.26	21.79	15.97
WATER	35.47	36.19	41.21	44.83	42.7	38.27
RMSD						
BOND LENGTHS (Å)	0.016	0.014	0.009	0.013	0.018	0.016
BOND ANGLES (˚)	1.455	1.371	1.042	1.319	1.456	1.368

Table A2.6. Gibbs energy profiles computed at B3LYP-D3BJ/6-311+G(d,p)+Fe(LanL2DZ)(PCM=Diethylether)//B3LYP/6-31G(d)+Fe(LanL2DZ), considering doublet and quadruplet electronic states. Energies are given in kcal/mol.

ΔG					
Path A	Reactants	TS1-a	Int1-a complex	TS2-a	Product 4 formation
doublet	0.2	18.6	*(barrierless)*		-55.7
quartet	0.0	20.8	12.6	15.1	-55.7

ΔG					
Path B	Reactants	TS1	int - complex	TS2	Product 7 formation
doublet	0.2	15.7	3.3	*(barrierless)*	-52.4
quartet	0.0	15.7	5.8	1.6	-52.4

ΔG					
Path C	Reactants	TS1	int - complex	TS2	Product 7 formation
doublet	0.2	4.3	(*could not be optimized*)	7.3	-52.4
quartet	0.0	3.2	-7.1	8.4	-52.4

Table A2.7. GcoA sequence conservation analysis.

resid (GcoA)	AA (GcoA)	relative entropy	z-score	percentile
1	M	3.28	2.59	98.0
2	T	1.42	-0.18	51.4
3	T	0.95	-0.89	16.5
4	T	1.44	-0.16	52.6
5	E	0.59	-1.42	1.7
6	R	0.63	-1.36	2.7
7	P	0.69	-1.28	4.7
8	D	0.79	-1.12	9.1
9	L	0.67	-1.31	3.7
10	A	1.02	-0.78	20.9
11	W	1.49	-0.09	55.5
12	L	0.96	-0.87	17.7
13	D	1.04	-0.75	21.6
14	E	1.10	-0.67	27.8
15	V	1.67	0.19	64.4
16	T	2.10	0.83	81.3
17	M	1.10	-0.67	28.0
18	T	0.84	-1.06	11.8
19	Q	1.28	-0.39	43.2
20	L	2.17	0.94	83.8
21	E	0.96	-0.87	17.4
22	R	0.84	-1.06	11.5
23	N	2.14	0.89	83.0
24	P	2.56	1.52	91.6
25	Y	3.15	2.40	97.3
26	E	1.95	0.61	77.6
27	V	1.30	-0.37	44.2
28	Y	3.24	2.54	97.5
29	E	1.05	-0.74	22.6
30	R	2.11	0.84	81.6
31	L	1.98	0.65	78.6
32	R	2.43	1.32	88.2
33	A	0.98	-0.85	18.9
34	E	1.81	0.39	71.5
35	A	1.20	-0.52	36.9
36	P	2.43	1.32	88.5
37	L	1.74	0.29	67.8
38	A	1.20	-0.51	37.3
39	F	1.64	0.15	62.9
40	V	1.79	0.36	70.0
41	P	1.61	0.10	61.4
42	V	0.88	-1.00	13.5
43	L	0.97	-0.85	18.7
44	G	1.62	0.11	62.2
45	S	0.75	-1.19	6.6
46	Y	1.65	0.16	63.6
47	V	1.38	-0.25	47.7
48	A	1.28	-0.39	43.5
49	S	2.13	0.88	82.6
50	T	1.53	-0.03	57.5
51	A	1.09	-0.68	26.8
52	E	1.19	-0.54	35.9
53	V	0.92	-0.93	15.5
54	C	2.79	1.86	95.6
55	R	1.06	-0.72	24.1
56	E	0.73	-1.22	6.1
57	V	1.71	0.25	66.1
58	A	1.07	-0.71	24.8

59	T	0.69	-1.28	4.2
60	S	1.37	-0.27	47.4
61	P	1.01	-0.80	19.9
62	D	1.06	-0.72	24.3
63	F	2.67	1.69	94.3
64	E	1.17	-0.56	34.2
65	A	1.30	-0.37	44.7
66	V	0.82	-1.08	10.6
67	I	0.96	-0.87	17.2
68	T	1.03	-0.77	21.4
69	P	1.02	-0.79	20.6
70	A	1.06	-0.72	23.6
71	G	1.25	-0.43	41.3
72	G	0.84	-1.06	11.3
73	R	1.87	0.48	74.9
74	T	1.27	-0.41	41.8
75	F	2.23	1.02	85.5
76	G	2.00	0.68	79.4
77	H	1.18	-0.55	34.6
78	P	1.61	0.10	61.7
79	A	1.32	-0.34	45.7
80	I	1.48	-0.10	54.8
81	I	1.81	0.40	71.7
82	G	1.12	-0.63	30.7
83	V	1.46	-0.12	53.8
84	N	1.84	0.44	73.2
85	G	2.06	0.77	80.6
86	D	1.29	-0.37	44.0
87	I	0.83	-1.07	10.8
88	H	3.61	3.09	99.3
89	A	0.76	-1.17	7.4
90	D	0.79	-1.12	9.3
91	L	1.44	-0.15	53.1
92	R	2.60	1.58	93.1
93	S	0.78	-1.15	8.1
94	M	1.18	-0.55	34.9
95	V	1.74	0.29	67.3
96	E	1.15	-0.60	31.9
97	P	1.69	0.22	65.1
98	A	0.85	-1.04	12.0
99	L	1.86	0.47	74.0
100	Q	1.38	-0.24	48.2
101	P	2.15	0.91	83.3
102	A	1.00	-0.81	19.7
103	E	0.74	-1.21	6.4
104	V	1.80	0.38	71.3
105	D	1.04	-0.75	21.9
106	R	0.72	-1.24	5.7
107	W	1.50	-0.07	56.5
108	I	1.58	0.06	59.7
109	D	0.80	-1.12	9.8
110	D	0.90	-0.96	14.7
111	L	1.39	-0.23	48.6
112	V	1.51	-0.04	57.0
113	R	1.42	-0.19	51.1
114	P	1.09	-0.67	27.5
115	I	1.40	-0.21	49.9
116	A	1.49	-0.08	55.8
117	R	1.09	-0.68	26.5
118	R	0.81	-1.10	10.1
119	Y	1.91	0.55	76.7
120	L	1.40	-0.21	50.1
121	E	0.97	-0.86	18.4

122	R	0.77	-1.16	7.9
123	F	1.65	0.16	63.9
124	E	1.10	-0.66	29.0
125	N	0.75	-1.18	6.9
126	D	0.88	-1.00	14.0
127	G	2.33	1.18	87.2
128	H	0.78	-1.14	8.8
129	A	1.11	-0.66	29.2
130	E	1.87	0.49	75.2
131	L	1.96	0.63	77.9
132	V	1.20	-0.52	36.4
133	A	0.65	-1.33	3.2
134	Q	1.11	-0.65	29.5
135	Y	2.23	1.03	85.7
136	C	2.19	0.97	84.5
137	E	1.43	-0.17	51.8
138	P	2.58	1.55	91.9
139	V	1.55	0.01	59.0
140	S	1.91	0.54	76.4
141	V	1.21	-0.49	38.3
142	R	1.61	0.10	60.9
143	S	1.32	-0.33	46.4
144	L	1.85	0.46	73.5
145	G	1.17	-0.56	34.4
146	D	1.23	-0.47	39.8
147	L	1.24	-0.45	40.5
148	L	1.78	0.36	69.8
149	G	2.62	1.60	93.6
150	L	1.74	0.29	67.6
151	Q	0.43	-1.66	0.2
152	E	1.22	-0.48	39.1
153	V	1.30	-0.37	44.5
154	D	1.21	-0.49	38.1
155	S	0.86	-1.02	12.5
156	D	1.09	-0.68	27.3
157	K	1.47	-0.10	54.3
158	L	1.93	0.58	77.1
159	R	1.32	-0.33	46.7
160	E	1.32	-0.34	45.9
161	W	3.88	3.50	99.8
162	F	2.38	1.26	87.7
163	A	1.89	0.52	76.2
164	K	0.72	-1.23	5.9
165	L	1.99	0.66	78.9
166	N	1.22	-0.48	39.3
167	R	0.71	-1.25	5.2
168	S	1.75	0.30	68.3
169	F	1.09	-0.68	27.0
170	T	1.30	-0.36	45.2
171	N	3.30	2.63	98.3
172	A	1.14	-0.61	31.0
173	A	0.82	-1.09	10.3
174	V	0.97	-0.86	17.9
175	D	1.54	0.00	58.5
176	E	1.64	0.15	62.7
177	N	1.93	0.57	76.9
178	G	2.62	1.60	93.9
179	E	1.83	0.44	73.0
180	F	3.02	2.21	96.3
181	A	1.41	-0.20	50.4
182	N	2.72	1.75	94.8
183	P	1.67	0.19	64.6
184	E	1.24	-0.46	40.0

185	G	0.84	-1.06	11.1
186	F	1.97	0.64	78.4
187	A	0.63	-1.37	2.5
188	E	0.66	-1.33	3.4
189	G	1.16	-0.58	33.2
190	D	1.82	0.41	72.2
191	Q	0.78	-1.14	8.4
192	A	1.48	-0.10	54.5
193	K	1.28	-0.40	42.8
194	A	0.94	-0.91	15.7
195	E	1.86	0.47	74.2
196	I	1.83	0.43	72.5
197	R	0.76	-1.17	7.1
198	A	0.94	-0.90	16.2
199	V	0.79	-1.12	9.6
200	V	1.40	-0.22	49.6
201	D	1.08	-0.69	25.8
202	P	1.03	-0.77	21.1
203	L	1.06	-0.72	23.8
204	I	1.16	-0.57	33.4
205	D	1.15	-0.59	32.4
206	K	1.14	-0.60	31.4
207	W	1.50	-0.07	56.8
208	I	0.86	-1.02	12.8
209	E	0.54	-1.51	0.7
210	H	1.16	-0.58	32.9
211	P	2.54	1.49	90.4
212	D	1.87	0.49	75.7
213	D	0.90	-0.96	15.0
214	S	2.21	1.00	85.0
215	A	1.10	-0.66	28.5
216	I	1.41	-0.20	50.6
217	S	2.84	1.94	95.8
218	H	1.82	0.41	72.0
219	W	2.23	1.03	86.0
220	L	1.38	-0.24	47.9
221	H	2.10	0.82	81.1
222	D	1.17	-0.56	33.9
223	G	1.47	-0.11	54.1
224	M	1.87	0.48	74.7
225	P	2.22	1.02	85.3
226	P	1.44	-0.16	52.8
227	G	2.48	1.40	89.2
228	Q	1.66	0.17	64.1
229	T	1.53	-0.01	58.2
230	R	1.58	0.06	60.0
231	D	1.15	-0.59	32.2
232	R	0.77	-1.16	7.6
233	E	1.19	-0.53	36.1
234	Y	1.24	-0.46	40.3
235	I	2.05	0.75	80.1
236	Y	1.70	0.23	65.4
237	P	1.59	0.07	60.4
238	T	2.13	0.88	82.3
239	I	1.55	0.00	58.7
240	Y	2.26	1.07	86.2
241	V	1.72	0.27	66.6
242	Y	1.14	-0.60	31.2
243	L	1.73	0.28	67.1
244	L	1.43	-0.17	51.6
245	G	2.44	1.35	88.9
246	A	1.80	0.38	71.0
247	M	1.87	0.49	75.4

248	Q	2.73	1.78	95.1
249	E	2.54	1.50	90.9
250	P	2.54	1.49	90.7
251	G	1.73	0.27	66.8
252	H	2.52	1.46	90.2
253	G	1.18	-0.54	35.1
254	M	1.28	-0.40	42.5
255	A	1.01	-0.79	20.1
256	S	1.79	0.36	70.3
257	T	1.43	-0.16	52.3
258	L	1.40	-0.22	49.1
259	V	1.08	-0.70	25.3
260	G	2.11	0.85	81.8
261	L	2.37	1.24	87.5
262	F	1.68	0.20	64.9
263	S	1.04	-0.75	22.1
264	R	1.61	0.11	61.9
265	P	2.19	0.97	84.3
266	E	1.57	0.03	59.2
267	Q	3.39	2.76	98.5
268	L	1.32	-0.34	46.2
269	E	1.16	-0.57	33.7
270	E	0.94	-0.91	16.0
271	V	2.06	0.77	80.3
272	V	0.89	-0.98	14.3
273	D	0.87	-1.01	13.0
274	D	1.53	-0.03	57.2
275	P	2.16	0.92	83.5
276	T	0.57	-1.45	1.0
277	L	1.12	-0.63	30.5
278	I	1.48	-0.09	55.3
279	P	1.10	-0.67	28.3
280	R	1.14	-0.60	31.7
281	A	1.88	0.50	75.9
282	I	1.76	0.32	68.8
283	A	1.05	-0.73	22.9
284	E	2.65	1.66	94.1
285	G	1.76	0.32	68.6
286	L	1.39	-0.24	48.4
287	R	2.60	1.58	92.9
288	W	3.80	3.37	99.5
289	T	1.32	-0.33	46.9
290	S	1.64	0.15	63.1
291	P	2.67	1.69	94.6
292	I	2.59	1.57	92.6
293	W	3.60	3.07	99.0
294	S	1.75	0.30	68.1
295	A	1.18	-0.54	35.4
296	T	1.28	-0.40	43.0
297	A	0.90	-0.96	14.5
298	R	2.38	1.26	88.0
299	I	0.96	-0.87	17.0
300	S	1.12	-0.64	30.2
301	T	1.48	-0.09	55.0
302	K	1.06	-0.73	23.1
303	P	1.53	-0.02	57.7
304	V	1.44	-0.15	53.3
305	T	1.39	-0.23	48.9
306	I	1.49	-0.08	56.0
307	A	0.97	-0.86	18.2
308	G	2.30	1.14	86.5
309	V	1.29	-0.38	43.7
310	D	0.48	-1.59	0.5

311	L	1.83	0.43	72.7
312	P	1.78	0.35	69.5
313	A	1.24	-0.45	41.0
314	G	2.31	1.15	87.0
315	T	1.02	-0.79	20.4
316	P	0.62	-1.38	2.2
317	V	2.05	0.75	79.9
318	M	1.22	-0.48	39.6
319	L	0.85	-1.03	12.3
320	S	1.18	-0.54	35.6
321	Y	1.77	0.33	69.0
322	G	1.60	0.08	60.7
323	S	2.43	1.33	88.7
324	A	2.19	0.97	84.8
325	N	2.97	2.14	96.1
326	H	2.13	0.88	82.1
327	D	2.49	1.41	89.4
328	T	1.26	-0.43	41.5
329	G	0.69	-1.28	4.4
330	K	0.78	-1.14	8.6
331	Y	2.51	1.44	89.7
332	E	0.92	-0.93	15.2
333	A	1.11	-0.65	29.7
334	P	1.77	0.33	69.3
335	S	1.10	-0.66	28.7
336	Q	0.61	-1.40	2.0
337	Y	2.79	1.86	95.3
338	D	1.99	0.67	79.1
339	L	1.41	-0.19	50.9
340	H	1.22	-0.49	38.8
341	R	2.60	1.58	93.4
342	P	0.99	-0.83	19.2
343	P	1.27	-0.41	42.0
344	L	0.69	-1.28	4.9
345	P	1.53	-0.02	58.0
346	H	3.27	2.58	97.8
347	L	1.43	-0.16	52.1
348	A	1.59	0.07	60.2
349	F	3.12	2.36	96.8
350	G	2.52	1.45	89.9
351	A	1.00	-0.81	19.4
352	G	2.31	1.14	86.7
353	N	1.07	-0.71	24.6
354	H	3.59	3.07	98.8
355	A	1.31	-0.34	45.5
356	C	4.79	4.86	100.0
357	A	1.40	-0.22	49.4
358	G	2.55	1.51	91.4
359	I	1.09	-0.68	26.3
360	Y	2.01	0.69	79.6
361	F	1.33	-0.31	47.2
362	A	1.65	0.16	63.4
363	N	1.71	0.24	65.8
364	H	0.88	-1.00	13.8
365	V	1.49	-0.08	56.3
366	M	1.24	-0.45	40.8
367	R	1.30	-0.36	45.0
368	I	1.95	0.61	77.4
369	A	1.20	-0.51	37.1
370	L	1.86	0.48	74.4
371	E	1.28	-0.40	42.3
372	E	1.21	-0.50	37.8
373	L	1.80	0.38	70.8

374	F	2.07	0.79	80.8
375	E	1.06	-0.73	23.3
376	A	1.05	-0.75	22.4
377	I	1.58	0.06	59.5
378	P	2.18	0.95	84.0
379	N	1.85	0.46	73.7
380	L	1.79	0.37	70.5
381	E	1.16	-0.58	32.7
382	R	1.21	-0.49	38.6
383	D	1.21	-0.50	37.6
384	T	0.87	-1.01	13.3
385	R	0.59	-1.43	1.5
386	E	0.68	-1.29	3.9
387	G	0.63	-1.36	2.9
388	V	1.20	-0.52	36.6
389	E	0.58	-1.44	1.2
390	F	1.61	0.10	61.2
391	W	1.46	-0.13	53.6
392	G	2.59	1.56	92.4
393	W	3.14	2.39	97.1
394	G	1.07	-0.71	25.1
395	F	3.03	2.23	96.6
396	R	2.55	1.51	91.2
397	G	1.97	0.63	78.1
398	P	1.72	0.26	66.3
399	T	1.08	-0.69	26.0
400	S	0.72	-1.24	5.4
401	L	2.14	0.89	82.8
402	H	1.63	0.12	62.4
403	V	1.70	0.24	65.6
404	T	0.95	-0.89	16.7
405	W	2.59	1.56	92.1
406	E	1.11	-0.64	30.0
407	V	1.08	-0.70	25.6

Table A2.8. *Pseudomonas putida* strains used in this study.

Strain ID	Genotype	Description of strain construction
CJ025	KT2440 Δ*catA2*	*catA2* was deleted from *P. putida* KT2440 using pCJ004 and this deletion was confirmed by diagnostic colony PCR amplification of a 2,089 bp product with primer pair oCJ084/oCJ085.
CJ028	KT2440 Δ*catBCA*::*Ptac*:*xylE* Δ*catA2*	*catBCA* was replaced in CJ025 with Ptac:*xylE* using pCJ005 and this gene replacement was confirmed by diagnostic colony PCR amplification of a 3,078 bp product with primer pair oCJ086/oCJ087.
AM140	KT2440 Δ*catBCA*::*Ptac*:*xylE* Δ*catA2* Δ*pcaHG*	*pcaHG* was deleted by transforming CJ028 with pCJ011. The gene replacement was confirmed by amplification of a 2049 bp fragment using primers oCJ106/oCJ107.
AM142	KT2440 Δ*catBCA*::*Ptac*:*xylE* Δ*catA2* Δ*pcaHG* / pBTL-*2*	AM140 transformed with pBTL-2
AM144	KT2440 Δ*catBCA*::*Ptac*:*xylE* Δ*catA2* Δ*pcaHG* / pCJ021 (*gcoAB* in pBTL-2)	AM140 transformed with pCJ021
AM148	KT2440 Δ*catBCA*::*Ptac*:*xylE* Δ*catA2* Δ*pcaHG* / p0AM27 ($gcoAB_{F169A}$ in pBTL-2)	AM140 transformed with pAM027
CJ612	KT2440 Δ*catBCA* Δ*catA2*	*catBCA* was deleted from CJ025 by transforming CJ025 with pCJ105. The gene deletion was confirmed by amplification of a 2407 bp product using primers oCJ086/oCJ087.
AM154	KT2440 Ptac:*pcaHG* Δ*catBCA* Δ*catA2*	Native promoter of *pcaHG* replaced with Ptac by transforming CJ612 with pCJ020. The gene replacement was confirmed by amplification of an 1201 bp fragment using primers oAM127/oCJ135.
AM155	KT2440 Ptac:*pcaHG* Δ*catBCA* Δ*catA2* / pBTL-2	AM154 transformed with pBTL-2
AM156	KT2440 Ptac:*pcaHG* Δ*catBCA* Δ*catA2* / pCJ021 (*gcoAB* in pBTL-2)	AM154 transformed with pCJ021
AM157	KT2440 Ptac:*pcaHG* Δ*catBCA* Δ*catA2* / ($gcoAB_{F169A}$ in pBTL-2)	AM154 transformed with pAM027

Table A2.9. Plasmids used in this study.

Plasmid ID	Use	Description of strain construction
pBTL-2	Empty vector	pBTL-2 was described previously in [503]
pCJ004	Deletion of *catA2* in *P. putida* KT2440	Construction of pCJ004 was described previously in Johnson and Beckham.[465]
pCJ005	Replacing *catBCA* with Ptac:*xylE* in *P. putida* KT2440	The 5' targeting region was amplified from *P. putida* KT2440 genomic DNA with primer pair oCJ042/oCJ043 (1,104 bp, which incorporated the tac promoter), *xylE* (969 bp) was amplified from *P. putida* mt-2 (ATCC 23973) genomic DNA with primer pair oCJ044/oCJ045, and the 3' targeting region was amplified using primer pair oCJ046/oCJ047 (1033 bp). These fragments were then assembled into pCM433 digested with AatII and SacI (7991 bp).
pCJ011	Deletion of *pcaHG* in *P. putida* KT2440	Construction of pCJ011 was described previously in Johnson and Beckham.[465]
pCJ020	Insertion of tac promoter upstream of *pcaHG* in *P. putida* KT2440	Construction of pCJ020 was described previously in Johnson and Beckham.[465]
pCJ021	Episomal expression of *gcoAB* on pBTL-2	Construction of pCJ021 was described previously in Tumen-Velasquez, *et al.*[282]
pCJ105	Deletion of *catBCA* in *P. putida* KT2440	The 5' (1054 bp) and 3' (1044 bp) targeting regions were amplifed from pCJ005 with primers oCJ542/oCJ543 and oCJ544/oCJ545, respectively, and assembled into pK18sB (Genbank MH166772 [504]) digested with EcoRI and HindII.
pAM027	Episomal expression of $gcoAB_{F169A}$ on pBTL-2	pCJ021 was linearly amplified using oAM173 and oAM174 to introduce the gcoA F169 mutation by site-directed mutagenesis and treated with NEB KLD Enzyme Mix (New England Biolabs), which includes kinase, ligase, and DpnI enzymes to phosphorylate and circularize the PCR product and digest the template, according to the manufacturer's instructions.
pGcoA-F196A, pGcoA-F196H, pGcoA-F196I, pGcoA-F196L, pGcoA-F196S, pGcoA-F196V	Expression of GcoA containing F196A, F196H, F196I, F196L, F196S, or F196V mutations	pCJ047 (pGEX-6P-1 vector containing WT $GcoA^{70}$) was linearly amplified using forward primers F196A, F196H, F196I, F196L, F196S, or F196V and F169rev to introduce the various gcoA mutations by site-directed mutagenesis and treated with NEB KLD Enzyme Mix (New England Biolabs), which includes kinase, ligase, and DpnI enzymes to phosphorylate and circularize the PCR product and digest the template, according to the manufacturer's instructions.

Table A2.10. Sequences of DNA oligos used in this study.

F169rev	AGCTTGGCGAACCACTCG
F169A	GAACCGCTCGgcCACCAACGCC
F169H	GAACCGCTCGcaCACCAACGCC
F169I	GAACCGCTCGaTCACCAACGC
F169L	GAACCGCTCGcTCACCAACGC
F169S	GAACCGCTCGTcCACCAACGC
F169V	GAACCGCTCGgTCACCAACGC
oCJ042	ccgaaaagtgccacctGACGTCcctgttgctcgatcaacgc
oCJ043	tcataAGATCTctcctgtgtgaaattgttatccgctcacaattccacacattatacgagccg atgattaattgtcaacagctctgttgccaggtcccgtc
oCJ044	aggagAGATCTtatgaacaaaggtgtaatgcgacc
oCJ045	cgaacGCGGCCGCgcaataagtcgtaccggaccatc
oCJ046	attgcGCGGCCGCgttcgaggttatgtcactgtgattttg
oCJ047	gctggatcctctagtGAGCTCcgcctgctccaggttg
oCJ084	CCTCAATGGCTTTGCCAG
oCJ085	GTACAACACACTGCCAGC
oCJ086	TGTGGGCATGGTGTGTTC
oCJ087	TCTTCAAAGCGTCCGGTG
oCJ106	ATCTTGAACCAACGCACC
oCJ107	CACAAGGCAATCCTGATCG
oCJ135	AGGCTGATGTTGATGTGC
oCJ542	aggaaacagctatgacatgattacGAATTCcctgttgctcgatcaacgccag
oCJ543	cgaacGCGGCCGCtgttgccaggtcccgtcagg
oCJ544	gacgggacctggcaacaGCGGCCGCgttcgag
oCJ545	cgttgtaaaacgacggccagtgccAAGCTTcgcctgctccaggttgaatgc
oAM127	GAGCTGTTGACAATTAATCATCGGC
oAM173	GAACCGCTCGgcCACCAACGCC
oAM174	AGCTTGGCGAACCACTCG

Table A.2.11 DFT optimized geometries

Electronic energies (E), zero point energy (ZPE), free energy (G(T)), quasiharmonic corrected free energy (qh-G(T)), and electronic energy from high level single point calculation (E Single point) of all stationary points (in a.u.). Cartesian coordinates are reported in xyz format.

Structure	E (au)	ZPE (au)	G(T) (au)	qh-G(T) (au)	E Single Point (au)
1 - syringol	-536.495040	0.170476	-536.361137	-536.360232	-536.706360
Fe=O-Porph - doublet	-1625.106262	0.317660	-1624.839694	-1624.837490	-1625.567736
Fe=O-Porph - quartet	-1625.106093	0.317696	-1624.840115	-1624.837883	-1625.567464
FeH2O-Porph - doublet	-1626.387773	0.338497	-1626.101953	-1626.099515	-1626.878338
3 - hemiacetal	-611.711120	0.175635	-611.573661	-611.572350	-611.959659
7 - acetal	-535.284576	0.149181	-535.170634	-535.169544	-535.479338
Fe-Porph - sextet	-1549.964112	0.312990	-1549.704649	-1549.700465	-1550.402389
Fe=O-Porph - doublet + Syringol (reactant complex)	-2161.610445	0.489091	-2161.193023	-2161.181629	-2162.291362
Fe=O-Porph - quartet + Syringol (reactant complex)	-2161.610252	0.489162	-2161.193304	-2161.181929	-2162.290915
TS1-a - doublet	-2161.574263	0.482285	-2161.159708	-2161.150955	-2162.264394
TS1-b - doublet	-2161.578885	0.482284	-2161.163139	-2161.155284	-2162.269456
TS1-a - quartet	-2161.572822	0.482653	-2161.160202	-2161.149881	-2162.260628
TS1-b - quartet	-2161.578118	0.482402	-2161.163372	-2161.155008	-2162.268868
Int1-b - doublet	-2161.595046	0.484420	-2161.180622	-2161.170366	-2162.290290
Int1-a - quartet	-2161.585020	0.485263	-2161.172652	-2161.160884	-2162.274849
Int1-b - quartet	-2161.593829	0.485312	-2161.178962	-2161.168993	-2162.286346
TS2-b - quartet	-2161.590346	0.480888	-2161.177325	-2161.169410	-2162.289260
TS2-a - quartet	-2161.579121	0.484609	-2161.165749	-2161.154995	-2162.270819
TS1-c - doublet	-2161.598186	0.483285	-2161.182677	-2161.174168	-2162.287930
TS1-c - quartet	-2161.599963	0.483375	-2161.184856	-2161.176470	-2162.289148
Int1-c - quartet	-2161.623252	0.487968	-2161.207527	-2161.195638	-2162.309777
TS2-c - doublet	-2161.593206	0.481997	-2161.180272	-2161.170068	-2162.282252
TS2-c - quartet	-2161.589202	0.481568	-2161.176370	-2161.167267	-2162.279422

A3 Supporting Information for Characterization of a two-enzyme system for plastics depolymerization.

A3.1 Supplementary Materials and Methods

A3.1.1 Plasmid construction

pET-21b(+) (EMD Millipore)-based plasmids for expression of the various *Ideonella sakaiensis* PETase and MHETase enzymes, as well as homologous, and mutant proteins were either synthesized by Twist Bioscience or constructed using NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs) and/or the Q5® Site-Directed Mutagenesis Kit (New England Biolabs) such that each protein has a C-terminal hexa-histidine epitope tag. For DNA assembly, DNA fragments were either amplified using Q5® High-Fidelity 2X Master Mix (New England Biolabs) or synthesized by Integrated DNA Technologies. Kits and master mixes were used according to the manufacturer's instructions. Plasmids were initially transformed into NEB® 5-alpha F'I^q Competent *E. coli* (New England Biolabs) and confirmed using Sanger sequencing by GENEWIZ, Inc.

A3.1.2 Protein expression and purification

For initial screening for soluble protein expression of the proteins of interest, various cell lines and induction methods were used to purify protein for kinetic assays.[505] For expression and purification, OverExpress™ *E. coli* C41 (DE3) (Lucigen) cells were transformed with pET21b(+) plasmid constructed with the gene of interest. Single colonies

from transformation were then inoculated into a starter culture of Luria Broth (LB) media containing 100 μg/mL ampicillin and grown at 37°C overnight. The starter culture was inoculated at a 100- fold dilution into a 2xYT medium containing 100 μg/mL ampicillin and grown at 37°C until the optical density measured at 600 nM (OD600) reached 0.6-0.8. Protein expression was then induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM. Cells were maintained at 20°C for 18 to 24 hours following IPTG induction, harvested by centrifugation, and stored at -80°C until purification. Harvested cells were resuspended in a lysis buffer (300 mM NaCl, 10 mM imidazole, 20 mM Tris HCl, pH 8.0,) and lysed using a bead beater (BioSpec Products, Inc.). Lysate was clarified by centrifugation at 40,000 x *g* for 45 minutes. Clarified lysate was then applied to a 5 mL HisTrap HP (GE Healthcare) Ni-NTA column using an ÄKTA Pure chromatography system (GE Healthcare) and eluted using 300 mM NaCl, 300 mM imidazole, 20 mM Tris HCl, pH 8.0. Resulting fractions containing proteins of interest were applied to a Sephacryl S-100 26/60 HR (GE Healthcare) size exclusion column equilibrated with 100 mM NaCl, 20 mM Tris HCl, pH 7.5 for biochemical assays, or the fractions were applied to a Superdex 75 pg 16/60 (GE Healthcare) size exclusion column equilibrated with 100 mM NaCl, 20 mM Tris HCl, pH 7.5 for crystallography. Protein in eluted fractions from Ni-NTA and size exclusion columns were assessed using SDS-PAGE with Coomassie staining and Western blot using primary antibody against the hexa-histidine epitope tag (Invitrogen). Total protein was assessed by BCA assay.[506] For proteins that did not express, or expressed in inclusion bodies, using the above described expression protocol, additional *E. coli* expression cell lines were tested, including Rosetta 2 (DE3) (Novagen), BL21 (DE3), and Lemo21 (DE3) (New England Biolabs), as was expression

by autoinduction at 30°C in ZYP-5052 media.[505] All studies reported were performed with freshly purified protein. The use of lyophilized protein was attempted, however, specifically for MHETase, inconsistent enzyme inhibition behaviors were observed.

Chimera proteins were expressed and purified as described above with the following noted exceptions: Single colonies from transformation into C41 (DE3) competent cells were used to inoculate a starter culture of 200 mL Terrific Broth (TB) media containing 100 µg/mL ampicillin for overnight outgrowth at 37°C. From the starter culture, 50 mL was used to inoculate 1 L of TB media containing 100 µg/mL ampicillin. For purification, cells were disrupted by sonication. In the final chromatography step a Superdex 200 pg 16/600 (GE Healthcare) size exclusion column equilibrated with 100 mM NaCl, 20 mM Tris HCl, pH 7.5 was used.

A3.1.3 Crystallography

After purification, as described above, MHETase protein was concentrated to a range of concentrations (9-14 mg/mL) and dialyzed into 100 mM NaCl, 10 mM Tris, pH 7.0 for crystallography. For seleno-methionine labeling of MHETase, K-MOPS minimal media was used.[507] Cells were grown to an OD_{600} of 0.5 after which 100 mg/L of DL-seleno-methionine (Sigma), 100 mg/L lysine, threonine and phenylalanine, leucine, isoleucine and valine were added as solids. IPTG (1 mM final concentration) was then added after 20 min and cells were grown for a further 16 h at 20°C. Seleno-methionine labeled protein was purified as described above. MHETase was crystallized at a range of concentrations from 9-14 mg/mL by sitting-drop vapor diffusion. Several conditions

yielded crystals, four of which were used to obtain datasets, one of which contained seleno-methionine labelled protein.

The crystals were cryo-cooled in liquid nitrogen after the addition of glycerol to 20% (v/v) while leaving the other components of the mother liquor at the same concentration. Seleno-methionine MHETase crystals belonging to space group $P22_12_1$ were used to obtain phase information using the I03 beamline at the Diamond Light Source (Oxford, UK). Data were obtained from 3600 images collected at 0.9795 Å with 0.1° increments. All images were integrated using XDS,[508] and scaled using SCALA.[509] Phases were obtained using PHASERSAD in the CCP4i software in combination with PARROT and SHELXD.[510, 511] The initial output was subsequently built using BUCCANEER and further refined using iterative rounds of COOT and PHENIX.[512-514] One molecule of MHETase was observed in the asymmetric unit of the $P22_12_1$ seleno-methionine SAD dataset. Three additional native datasets, each containing 1800 images collected at 0.1° increments, were collected at beamline I03 of the Diamond Light Source. The structure of native MHETase were obtained using molecular replacement from a refined molecule of MHETase obtained initially from the seleno-methionine SAD data. All structures were refined using iterative rounds of COOT and PHENIX.[512-514] Cell constants, crystallographic data and details of the refined models are summarized in Table A3.1.[515] Structural figures were generated with PYMOL (Schrödinger, LLC) with accompanying sequence alignments generated in Clustal W,[516] and rendered using ESPript 3.0.[517]

A3.1.4 Ligand synthesis

MHET, MHEI, and MHEF synthesis. Mono(2-hydroxyethyl) terephthalate (MHET) and mono(2-hydroxyethyl) isophthalate (MHEI) and mono(2-hydroxyethyl) furanoate (MHEF) were synthesized via the condensation of either terephthloyl chloride, isophthloyl chloride, or the acyl chloride of 2,5-furan dicarboxylic acid, respectively, with monoprotected ethylene glycol (tBOC-EG) which was subsequently deprotected to yield the final product.

Initially, tBOC-EG was prepared by stirring one molar equivalent of ethylene glycol (EG) with one molar equivalent of di-tert-butyl decarbonate with 0.01 equivalents of 4-dimethylaminopyridine (DMAP) as a catalyst in dichloromethane (DCM). The reaction mixture was allowed to stir for 24 hours and was subsequently washed with DI water, 1 M HCl, and brine follow by drying with sodium sulfate. The solvent was removed, and the product was purified via silica gel column chromatography to yield the mono-protected tBOC-EG. The yield of this reaction was 60% at a final purity of 99+% (via NMR and HPLC).

To form MHET, MHEI, or MHEF either one molar equivalent of terephthloyl chloride, isophthloyl chloride, or the acyl chloride of 2,5-furan dicarboxylic acid, respectively, was dissolved in DCM with one molar equivalent of tBOC-EG. One molar equivalent of triethylamine (TEA) was then added in dropwise over a period of 30 minutes. The reaction solution was subsequently washed with DI water and brine and then dried over sodium sulfate prior to removing the DCM. The crude product was subsequently taken up in a mixture of 10% acetone in DCM and purified via silica gel chromatography.
NMR for MHET, MHEF, and MHEI is provided in Figures A3.18, A3.19, and A3.20, respectively.

A3.1.5 Enzyme activity and synergy with PETase

Quenching enzymatic reactions. Previous studies of MHETase activity report the use of an equal volume of pH 2.5 sodium phosphate buffer and a heat treatment at 80-85°C for 10 min to quench enzymatic activity.[85, 518] We found this to be an inconsistent method of quenching enzyme activity, such that some level of enzyme activity continues after treatment, as quantified by HPLC analysis. To determine a reliable method for quenching enzymatic activity we performed quenching trial experiments in triplicate for reactions containing 250 µM MHET, 90 mM NaCl, 10% (v/v) DMSO, 45 mM sodium phosphate, pH 7.5, at 30°C, for both reactions containing enzyme in order to compare the quenching capacity of a given method, and reactions without enzyme to evaluate the level of non-enzymatic hydrolysis caused by the treatment method. Quenching solution components intended to denature the enzyme, such as a reducing agent (TCEP), chaotropic agent (GuHCl), or strong acid (6N HCl) proved either inadequate to completely quench enzymatic activity, or rather resulted in high levels of acid-mediated hydrolysis of the substrate. The active-site inhibitor PMSF inconsistently quenched enzymatic activity. Polar solvents (ethanol, methanol, and Isopropanol) were most effective at quenching enzymatic activity. The quenching solutions used are summarized in Table A3.6. Based on the results, an equal volume of 100% methanol followed by a heat treatment at 85°C for 10 min was selected as the most reliable method of quenching, which also yields the lowest levels of non-enzymatic hydrolysis of MHET.

Determination of enzyme turnover rates. Comparative assays for each enzyme were performed at the same enzyme and substrate concentration. Reactions were performed in

triplicate over a 15 min time course using 5 nM enzyme concentration and 250 μM MHET in 90 mM NaCl, 10% (v/v) DMSO, 45 mM sodium phosphate, pH 7.5, at 30°C. Reactions were terminated using an equal volume of 100% methanol followed by heat treatment at 85°C for 10 min. Product and substrate were quantified by HPLC. Apparent turnover rate (k_{cat}) was determined by terephthalic acid (TPA) produced.

Michaelis-Menten kinetics of MHETase and variants. Reactions were performed in triplicate over a 10 min time course using 5 nM enzyme and substrate concentrations ranging from 10 μM to 250 μM MHET in 90 mM NaCl, 10% (v/v) DMSO, 45 mM sodium phosphate, pH 7.5, at 30°C. Each reaction was terminated using an equal volume of 100% methanol and heat treatment at 85°C for 10 min. Product and substrate were quantified by HPLC. Initial reaction velocities were calculated from TPA produced over time and kinetic parameters were determined by nonlinear regression of the initial velocities fit to the Michaelis-Menten equation. The wild-type MHETase and both homologous enzymes were fitted to the Michaelis-Menten model with substrate inhibition (Equation A3.1) while the MHETase S131G mutant was fitted to the simple Michaelis-Menten model (Eq. 2) using GraphPad Prism version 8.4.1 for MacOS (GraphPad Software, San Diego, California USA), as follows:

$$v = \frac{V_{max}\,[S]}{K_m + [S](1 + \frac{[S]}{K_i})} \tag{A3.1}$$

$$v = \frac{V_{max}\,[S]}{K_m + [S]} \tag{A3.2}$$

While both substrate inhibition and product inhibition are possible in these reactions, the relationship between initial reaction velocity and initial substrate concentration indicates substrate inhibition predominates in these reaction conditions. Low

substrate concentrations were considered in these kinetic studies in order to minimize the effect of substrate inhibition.

Enzymatic degradation of PET film. Amorphous PET film (2-3% crystallinity, Goodfellow, UK) was incubated with enzyme of interest in polypropylene tubes containing 90 mM NaCl, 10% (v/v) DMSO, 45 mM sodium phosphate, pH 7.5, at 30°C for 96 hours. The reaction was terminated by addition of equal volume 100% methanol and PET coupons were removed from the reaction solution. The reaction solution was heat treated at 85°C for 10 minutes. PET coupons were washed with consecutive rinses of 1% SDS, 100% DMSO, ultrapure water, and 95% ethanol. Coupons were then vacuum dried for 24 h at 60°C for scanning electron microscopy.

Activity assay of MHETase with non-MHET substrates. Evaluation of MHETase activity was performed in triplicate using 5 nM enzyme concentration and 25 µM, 50 µM, and 250 µM substrate concentration at 30°C for 24 h in a 0.5 mL reaction volume. The reaction was carried out in 90 mM NaCl, 10% (v/v) DMSO, 45 mM sodium phosphate, pH 7.5, reaction buffer with three concentrations of each substrate (MHET, MHEI, or MHEF). Reactions commenced upon addition of enzyme or an equal volume of reaction buffer for the no enzyme controls. At the end of 24 h the reactions were terminated using an equal volume of 100% DMSO and heat treatment at 85°C for 10 min. Product and substrate were analyzed by HPLC. Values reported as percentage of substrate hydrolyzed into product.

HPLC method. Standards of BHET, TPA, 2,5-furandicarboxylic acid, and isophthalate were obtained from Sigma Aldrich. MHET, MHEI, and MHEF were synthesized as described above. Analyte analysis of samples was performed on an Agilent 1260 LC system (Agilent Technologies, Santa Clara, CA) equipped with a G1315A diode

array detector (DAD). Each sample and standard were injected using a volume of 10 µL onto a Phenomenex Luna C18(2) column, 5 µm, 4.6 x 150 mm (Phenomenex, Torrance, CA). The column temperature was maintained at 40°C and the mobile phase used to separate the analytes of interest was composed of 20 mM phosphoric acid in water (A) and 100% methanol (B). The separation was carried out using a constant flow rate of 0.6 mL/min and a gradient program of: at t = 0 min (A) = 80% and (B) = 20%; at t = 15 min (A) = 35% and (B) = 65%; at t = 15.01 min through 20 min (A) = 80% and (B) = 20% for a total run time of 20 min. The calibration curve for each analyte was evaluated between concentrations of 0.1 – 200 mg/L. DAD detection at a wavelength of 240 nm was performed for each analyte. Ten calibration standards were used with an r^2 coefficient of 0.995 or better and a calibration verification standard (CVS) at 100 mg/L for each analyte was analyzed every 18 samples to ensure the integrity of the initial calibration. Samples were diluted with an equal volume of ultrapure water for analysis.

A3.1.6 Scanning Electron Microscopy.

Dried PET coupons sized 2.5 cm x 0.5 cm were placed on aluminum stubs using carbon tape, and were sputter coated with 9 nm of iridium. SEM imaging was performed using an FEI quanta 400 FEG instrument under low vacuum (0.45 torr), beam-accelerating voltage of 25 keV.

A3.1.7 Bioinformatics

Sequence selection and conservation analysis. 6,671 tannase family sequences were retrieved by a PSI-BLAST search against the NCBI non-redundant database with *Is*

MHETase (A0A0K8P8E7.1) as initial query sequence on November 16, 2018. A total of three iterations of the PSI-BLAST search were carried out, and all 6,671 hits had E-values of 1e-50 or better. A multiple sequence alignment of the 6,671 tannase family sequences was carried out with MAFFT [519]. The amino acid conservation at each site of the multiple sequence alignment was evaluated by computing the relative entropy according to the following equation:[520]

$$R.E = \sum_{i=1}^{20} \left(p_i \log \frac{p_i}{p_i^{MSA}} \right) \quad \text{(A3.3)}$$

where p_i is the frequency of the i^{th} amino acid in the given site and p_i^{MSA} is the overall frequency of the i^{th} amino acid in the multiple sequence alignment.

Phylogenetic analysis. Through a keyword search of the NCBI protein database (https://www.ncbi.nlm.nih.gov/protein) with BioPython,[521] functional annotation for the 6,671 sequences was retrieved. From the sequence description in the NCBI database, 338 and 51 sequences of the 6,671 sequences were clearly annotated as ferulic acid esterases, or as tannases, respectively. Profile hidden Markov models (HMMs) were constructed for ferulic acid esterases and tannases with the dataset of 338 and 51 sequences, respectively, using the HMMER software (version 3.1b2).[522] Sequence identity thresholds of 95% and 60% were, respectively, applied to the set of 338 ferulic acid esterases and 51 tannases resulting in a set of 120 sequences (86 FAEs, 31 tannases, *Is*-MHETase, and 2 *Is*-MHETase close homologs). The 120 sequences were aligned with MAFFT[519] and phylogenetic analysis with 1000 bootstrap replicates was conducted in MEGA7.[523] For the phylogenetic tree, the evolutionary distances were computed using the JTT matrix-based method.[524] The

minimum evolution tree was searched using the Close-Neighbor-Interchange (CNI) algorithm[172] at a search level of 1, and the Neighbor-joining algorithm[525] was used to generate the initial tree. Gaps in the alignment were handled using pairwise deletion. There were a total of 1440 positions in the final dataset.

Identification of homologous enzymes. MHETase shares low sequence similarity (<53%) with most sequences in the tannase family, with the exception of two homologous sequences, one from *Comamonas thiooxydans* (strain: NCBI:txid363952, protein: Genbank WP_080747404.1) (24) and one from *Hydrogenophaga* sp. PML113 (strain: NCBI:txid1899350, protein:Genbank WP_083293388.1). In the time since identification of this *C. thiooxydans* sequence, this accession entry was removed from Genbank upon request of the submitter. Three other strains of *C. thiooxydans* also carry this sequence (protein:Genbank WP_034389536.1), though lacking 28 residues at the N-terminus. These include *C. thiooxydans* strains DS1 (protein:INSDC KGH18114.1), DF1 (protein:INSDC KGH28153.1), and DF2 (protein:INSDC DGH05124.1). Using the original protein accession sequence (WP_080747404.1), SignalP prediction indicates the sequence encodes a 70-residue signal sequence. We attempted expression of both *C. thiooxydans* and *Hydrogenophaga* sp. PML113 enzymes without the predicted signal peptide, however the enzymes did not express.

A3.1.8 Molecular docking

MHETase structure preparation. MHETase structure was taken from starting structures used for molecular dynamics simulations. The MHETase structure was prepared with Schrodinger's Protein Preparation Wizard in Schrodinger).[526-528] PropKa was used to

optimize hydrogen bonds at pH 7.0; OPLS3 force field[529] was used to conduct a restrained minimization on all heavy atoms (to ensure less than 0.30 Angstrom deviation from starting structure position).

Ligand structure preparation. MHET and MHEI structures were built in Schrodinger using Maestro Workspace tools. All ligands were energetically minimized using Schrodinger LigPrep,[530] according to OPLS3 force field.[531] Ionization states of MHET and MHEI were requested with Epik at pH of 7.0, [532, 533] although no additional ionization states were generated.

Flexible ligand/flexible receptor docking. Induced Fit Docking (IFD) is Schrodinger's flexible ligand/flexible receptor docking tool.[530, 532, 534-536] IFD utilizes two other Schrodinger modules, Prime for amino-acid side chain prediction and refinement, and Glide for ligand docking, to achieve binding site flexibility during docking simulations. Ligands were docked into MHETase active site (determined by co-crystallization with benzoic acid) by trimming (mutating and back-mutating) all residues within 5 Å of the catalytic triad, except for the catalytic triad. This was necessary as attempts to mutate catalytic triad residues to Alanine then back-mutate after initial docking (as is procedure in IFD) would result in chemically incompetent catalytic triad residues. After docking and amino-acid refinement, binding modes were scored and ranked using the Glide XP scoring function. Resultant predicted binding poses were then analyzed to determine if each pose would result in cleavage of an ester bond, such poses were determined to be chemically relevant. Those chemically relevant poses with the lowest predicted binding free energies (i.e. lowest Glide XP score) are discussed in detail in the Results.

A3.1.9 Molecular simulations

The starting point for molecular dynamics (MD) simulations was chain A of the 1.8 Å resolution structure (PDB code 6QZ4). This structure was chosen because it has electron density for the widest range of residues (6QZ3 lacks residues 36-39 and 6QZ1 lacks residues 56-60). The bound calcium ion and the crystal waters are maintained (sulfate is deleted). For a variety of residues with alternate conformations, conformation A was chosen for the following: Ser143, Ile149, Ser240, and Asn403. Conformation B was chosen for Ser401 and Leu486. Initial proposal for protonation states was given by H++ server (http://biophysics.cs.vt.edu/H++) at pH 7.0,[537] consistent with Yoshida *et al.* reaction conditions.[518] Of the acidic residues (glutamic and aspartic acid), Glu230 was determined to be protonated. For histidine residues, His91 and His528 are singly protonated at ND1, His293, His 467, and His488 are singly protonated at NE2, and His166 and His241 are doubly protonated. The overall charge on MHETase with these protonation states is -6; 6 sodium ions were added to the solution phase to neutralize. Five disulfides are formed: Cys51 - Cys92, Cys224 - Cys529, Cys303 - Cys320, Cys340 - Cys348, and Cys577 - Cys599.

All simulations were built using CHARMM version 43a1,[329] and simulated with the CHARMM36 force field for the protein,[538] CHARMM force field for carbohydrates, [539, 540] and TIP3P water molecules.[541] Topologies and forcefield parameters for MHET were generated by CGenFF program version 2.2.0, [542, 543] for use with CGenFF forcefield version 4.0.[544, 545]

The simulation box is cubic, with each box side approximately 110 Å long. Approximately 132,000 atoms are modeled in each system. Classical MD simulations of

150 ns in length were run in triplicate for the following five scenarios: 1), 2) free MHETase (no substrate bound) with calcium ion bound at calcium binding site (Phe415 open and closed), 3), 4) MHETase with MHET bound at active site with calcium ion bound at calcium binding site (Phe415 open and closed), and 5) free MHETase with neither substrate nor calcium ion bound (Phe415 open). For the simulations with bound MHET, the initial state was prepared as follows. Near neutral pH, MHET exists in solution as a salt, thus the carboxylate moiety of MHET is deprotonated in our simulations (for reference, the pKa of the first and second acidic moieties of TPA are 3.54 and 4.46 at 25°C (PubChem)). The initial configuration for MHET bound at the active site of MHETase was prepared in the following manner. *Is* PETase, bound with PET tetramer from a prior molecular docking study,[546] was trimmed back to a hydroxyethyl-capped PET dimer maintaining the ester bond nearest to the catalytic triad as well as the repeat units on either side. Following MM and QM/MM minimization**,** restraints were placed on two distances in order to prepare an enzyme-substrate configuration primed for catalysis: the nucleophilic attack distance between Ser225 oxygen and PET carbon (target: 2.0 Å), and the scissile C-O ester bond distance (target: 1.4 Å). Force constants of 200 kcal/mol/Å^2 were utilized in both cases. The catalytic residues of MHETase were then aligned with those of PETase. Subsequently trimming the PET dimer back to the heavy atoms it shares in common with MHET gave the starting point for MD simulations with MHET bound.

All classical MD simulations were performed at 303 K to match the conditions for hydrolytic assays performed by Yoshida *et al.*[518] Systems were density equilibrated for 1 ns at a constant pressure of 1 atmosphere and constant temperature of 303 K (controlled via the Nosé-Hoover barostat and thermostat); subsequent production runs were performed

with constant volume and temperature (303 K) in NAMD 2.9.[330] All bonded hydrogen distances were constrained utilizing the SHAKE algorithm.[547] The timestep was 2 fs. A nonbonded cutoff distance of 10 Å was utilized, with a switching distance of 9 Å, and a nonbonded pair list distance of 13 Å. The long-range electrostatics were described via the Particle Mesh Ewald (PME) method with a sixth order b-spline, a Gaussian distribution with a width of 0.312 Å, and 1 Å grid spacing. The velocity Verlet multiple timestepping integration scheme was used, with the full nonbonded interactions evaluated every timestep, full electrostatics interactions evaluated every 3 timesteps, and 6 timesteps between atom reassignments.

Following 1 ns of dynamics with classical forcefield, the CHAMBER utility[548] of ParmEd version 3.0.3 was used to convert the CHARMM coordinate, topology, parameter, and protein structure files to AMBER formatted coordinate and topology files for hybrid quantum mechanics/molecular mechanics (QM/MM) simulations by the sander program of AMBER version 12.[549] The AMBER software was used to carry out all QM/MM calculations,[360, 550] with the Self-Consistent Charge Density-Functional Tight-Binding (SCC-DFTB) semiempirical QM method using the Third-Order Parameterization for Organic and Biological Systems (3OB) to describe the QM region.[333] An 8 Å cutoff was used for nonbonded interactions and PME used for long-range electrostatics. Periodic boundary conditions were utilized, the timestep was 1 fs, and SHAKE was applied only to hydrogen atoms in the MM region (hydrogen atoms in the QM region were not constrained by the SHAKE algorithm). The Langevin thermostat and barostat were utilized with collision frequency of 1.0 ps^{-1} and pressure relaxation time of 2.0 ps.

The QM region includes the MHET substrate and the three catalytic residues (Ser225, Asp492, and His528, each cut across the C_α/C_β bond). For step 1 of the catalytic mechanism (acylation), there are 46 atoms in the QM region, with a QM region charge of -2. Hydrogen link atoms are utilized where covalent bonds cross the boundary between the QM and MM regions. For step 2 of the catalytic mechanism (deacylation), the ethylene glycol product is removed (as it was shown to leave the active site after the acylation step) and a single water molecule is added to the QM region, giving 39 atoms in the QM regions still with a charge of -2.

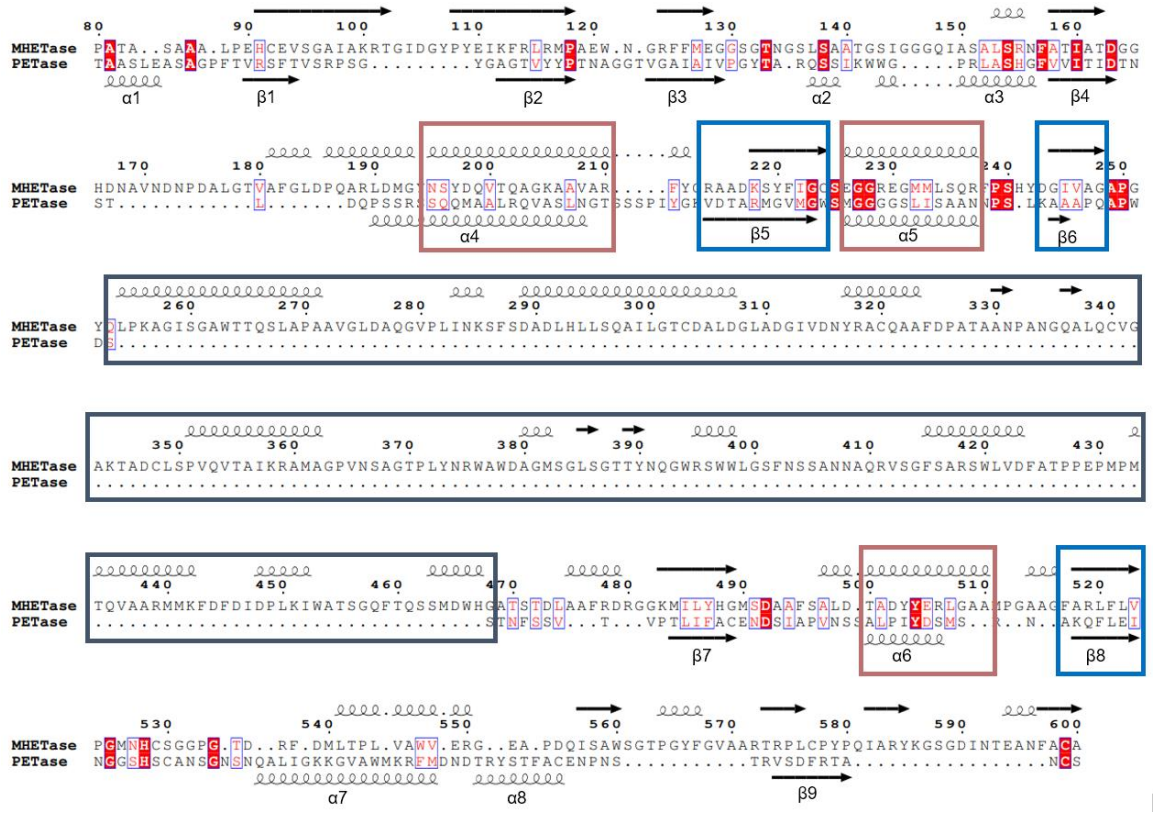
Two-dimensional free energy surfaces were prepared for acylation and deacylation via umbrella sampling simulations of each step. Order parameters utilized as reaction coordinates for both steps are the breaking and forming C-O bonds. For acylation, this is the breaking MHET ester bond ("r1") and the forming AEI bond between S125 and MHET carbonyl carbon ("r2"). For deacylation, this is the forming bond between water oxygen ("r1") and MHET carbonyl carbon and the breaking AEI bond between TPA and S125 ("r2"). Harmonic restraints are placed on these distances with force constant of 400 (kcal/mol)/Å^2. Windows are spaced by 0.1 Å increments for each of the two bonds (acylation: r1 between 1.3 and 3.4; r2 between 1.3 and 2.9; deacylation: r1 between 1.3 and 3.4; r2 between 1.3 and 3.9, neglecting some combinations that are particularly high energy in each case). Umbrella sampling simulations are performed in the NVE ensemble. Each window is equilibrated for 25 ps; data is collected on subsequent 500 ps. The variational free energy profile (vFEP) method[551] was utilized to produce the two-dimensional free energy profile from the probability distributions of each window. Block averaging (10

blocks for each set of umbrella sampling data) was utilized to estimate error bars on the free energy differences and reaction rate constants.

To aid in the visualization of the acylation and deacylation reactions in *Is* MHETase, QM/MM transition path sampling (TPS) simulations were undertaken.[552] In particular, the Aimless Shooting (AS) flavor of TPS was employed.[553, 554] TPS is a powerful technique for studying rare events because the trajectories that it generates are completely unrestrained and do not bias the reaction along any chosen reaction coordinate.[552] The AS variety has only one adjustable parameter (dt, which here is equal to 25 fs). The simulation time for each MHETase AS trajectory is 2 ps, which was sufficient for the trajectory to relax to both stable basins for reactant and product. Other simulation parameters, including the QM region, forcefield, timestep, cutoff distances, etc. are the same as in the QM/MM two-dimensional umbrella sampling simulations, described above. Path sampling simulations were undertaken purely for illustrative purposes; no data was analyzed from these trajectories. For both acylation and deacylation, the initial configuration (which represents a configuration that is putatively part of the transition state ensemble) was taken from the end point configuration of the US window restrained to distances of 1.9 Å for both the breaking and forming C-O bonds. For acylation, 6 out of 21 trajectories were accepted (meaning they connected reactants to products). For deacylation, 13 out of 23 trajectories were accepted. For both acylation and deacylation, the movie was made from the last accepted trajectory.

A3.2 Supplementary Figures

A



B

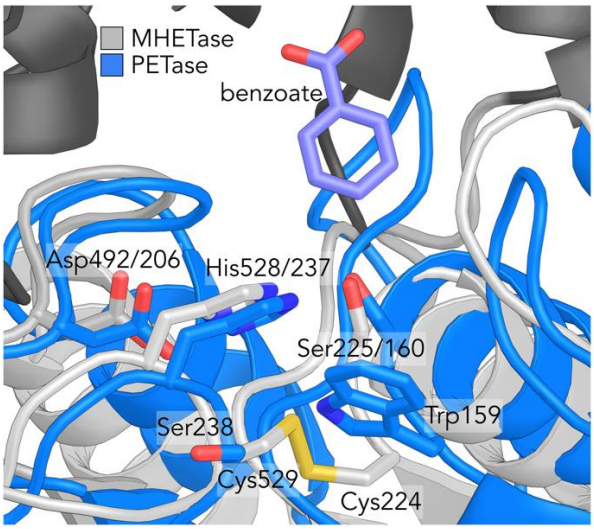


Figure A3.1: (A) Sequence alignment and secondary structure homology. The core sequence of *Ideonella sakaiensis* MHETase (residues 80-600) is shown aligned to PETase with regions of strong structural homology boxed in pink for common α-helices and boxed in blue for common β-strands. The lid domain is boxed in dark grey. The secondary structure elements are labelled according to the standard /β hydrolase nomenclature with α-helices depicted as spirals, β-strands as arrows, and numbering corresponding to the 3-dimensional representation presented in Figure A3.1D. Similar residues are shown in red text with identical residues in solid red boxes. B) Active site comparison of MHETase (PDB code 6QZ3) and PETase (PDB code 6EQE). Where two residues indices are given separated by a backslash, the first number applies to MHETase and the second to PETase. Lid domain of MHETase is shown in dark gray.

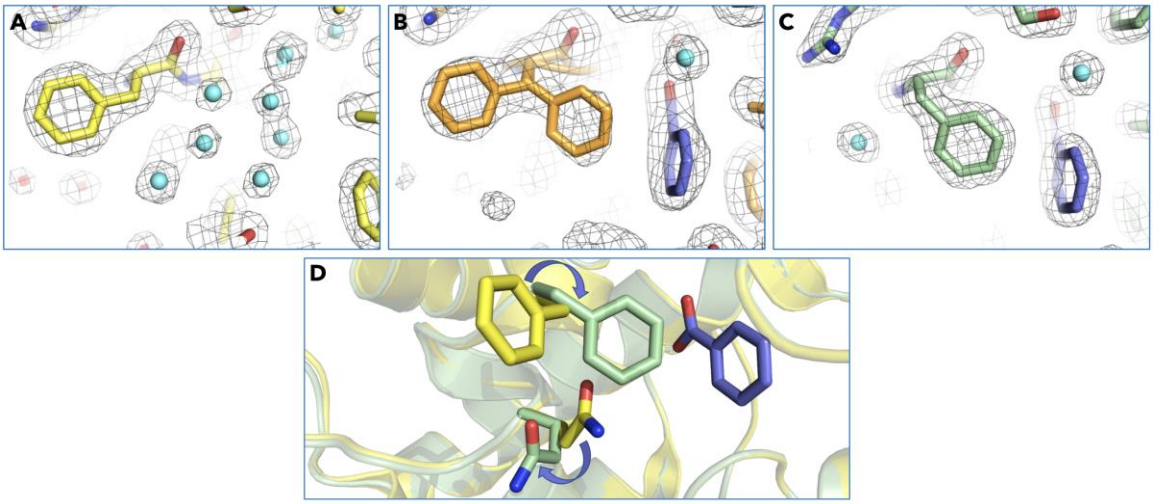


Figure A3.2: Alternate positions of residue Phe415 captured in multiple crystal structures. (A) The apo-structure (PDB ID: 6QZ4) with Phe415 depicted in yellow in the open conformation. The active site is populated with several water molecules (cyan spheres). The $2F_o–F_c$ electron density map was contoured at 1.3 σ. (B) A mixed conformation of Phe451 (orange) was refined in structure PDB ID: 6QZ1. Electron density for the benzoic acid (purple) was weaker than the surrounding residues, suggesting that the site it not fully occupied; hence the alternative positions shown here likely represent a mixture of bound and free states. The $2F_o–F_c$ electron density map was contoured at 0.5 σ to highlight the dual conformation. (C) The fully bound form of benzoic acid in the active site (PDB ID: 6QZ3) reveals Phe415 (green) in the closed conformation. The $2F_o–F_c$ electron density map was contoured at 1.3 σ. (D) As a point of reference to **Figure A3.1C**, the concerted movement of residues Gln410 and Phe415 on ligand binding is illustrated with purple arrows in a superposition of the apo enzyme (yellow) with the ligand bound state (green). The relative position of benzoic acid is depicted in purple.

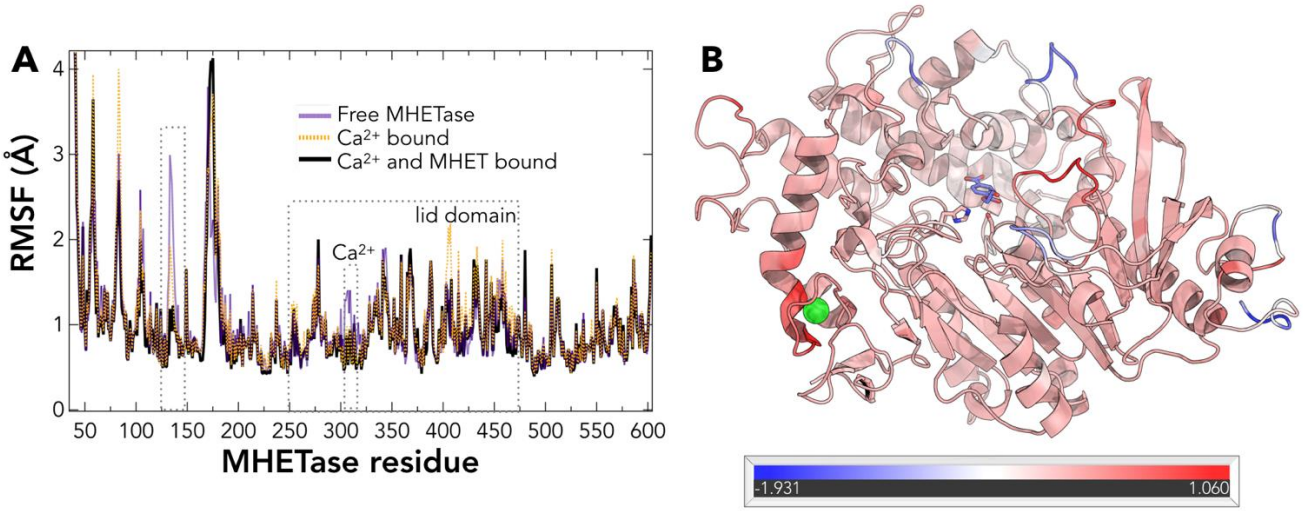


Figure A3.3: Effect of calcium binding on MHETase motion from molecular dynamics simulations. A) Root mean square fluctuations (RMSF) for the heavy atoms of each MHETase amino acid (backbone and side chain atoms) for three different situations. "Free MHETase" refers to MHETase bound with *neither* calcium ion nor MHET. "Ca^{2+} bound" has calcium bound at the calcium binding site but with empty active site. "Ca^{2+} and MHET bound" has calcium bound at calcium binding site and MHET bound at the active site. Each trace represents the average RMSF from three independent MD trajectories, each of 150 ns in length. RMSF analysis was performed in CHARMM. Shown in dashed boxes are the lid domain, the region immediately surrounding the Ca^{2+} binding site, and the loop region near the active site that is significantly stabilized by Ca^{2+} binding (approximately residues 125 through 150 and appearing in red in panel B). B) MHETase structure colored by the RMSF difference between "Free MHETase" (purple trace in panel A) and "Ca^{2+} bound" (orange trace in panel A) showing the regions wherein Ca^{2+} significantly stabilizes the enzyme (red), regions where there is little effect (white/pink), and those regions where the trend is actually reversed (blue).

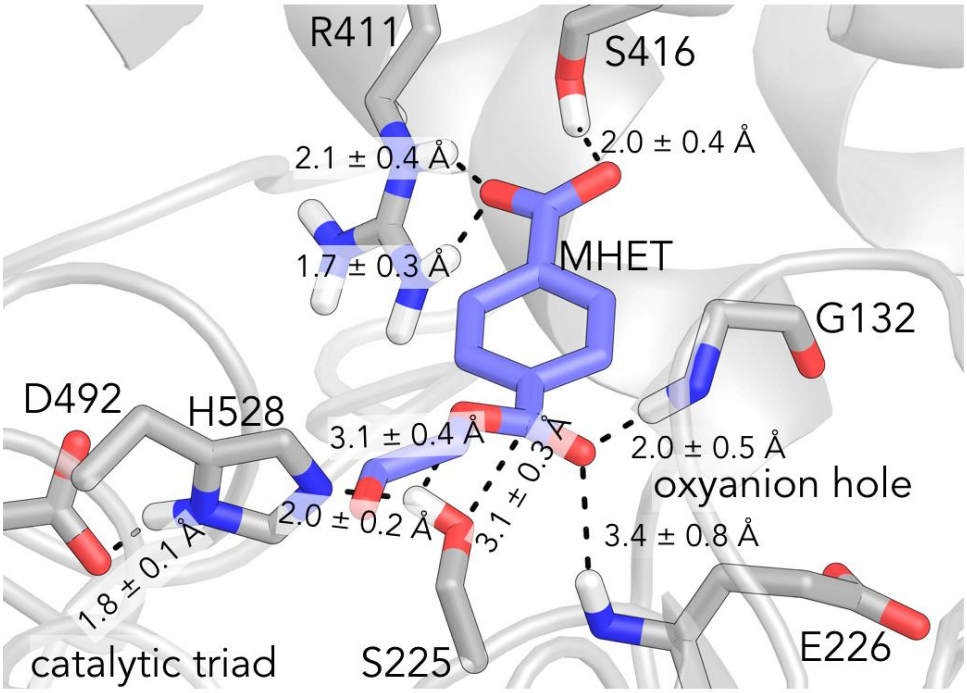


Figure A3.4: Molecular dynamics of MHET binding at MHETase active site. The distances noted represent the average ± standard deviation from three independent MD simulations, each of 150 ns in length. Gly132 and Glu226 comprise the oxyanion hole and interact with the carbonyl oxygen in the Michaelis complex and throughout the acylation reaction; Arg411 and Ser416 interact strongly with the carboxylate motif.

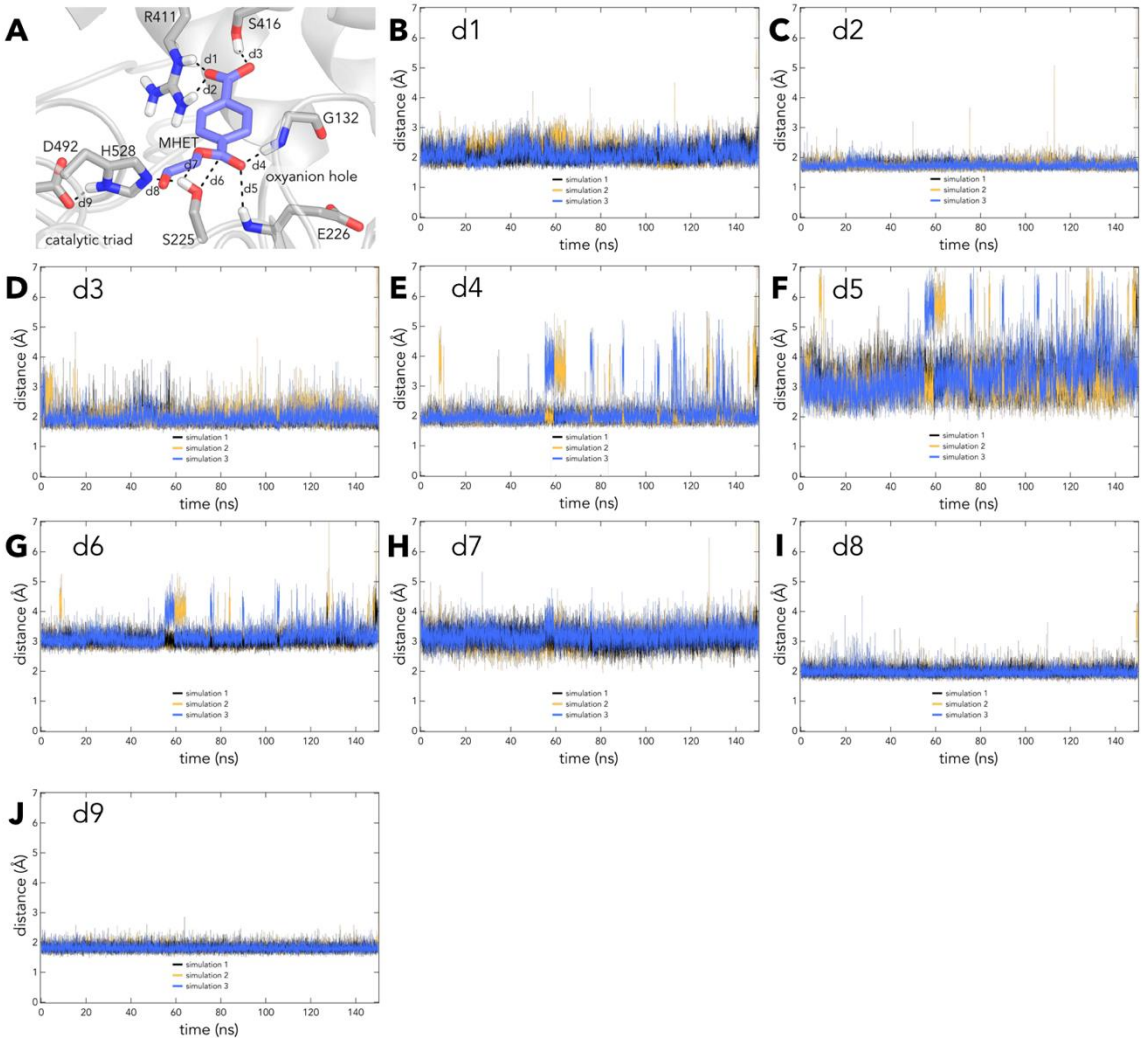


Figure A3.5: Time-traces for key distances in MD simulations of MHET bound at MHETase active site. Panels B-J show the dynamic time traces for the distances annotated in panel A. These are the same distances for which averages and standard deviations are shown in Figure A3.4. The three simulations referenced in each panel are identical MD simulations of 150 ns in length.

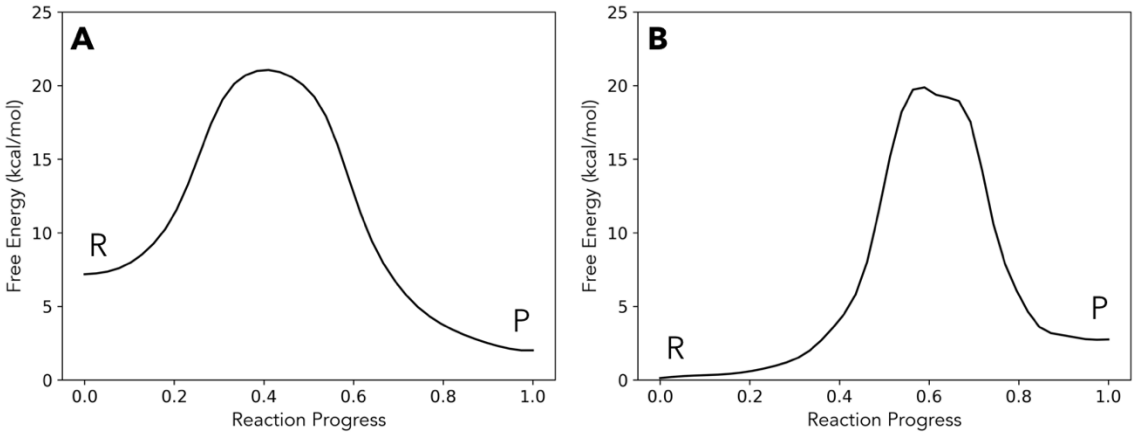


Figure A3.6: One-dimensional potentials of mean force (PMF) for acylation and deacylation steps. PMFs along the minimum free energy path (MFEP) for A) acylation reaction and B) deacylation reaction. The MFEPs were computed from the two-dimensional free energy surfaces. These 1D PMFs represent the free energy along the MFEP.

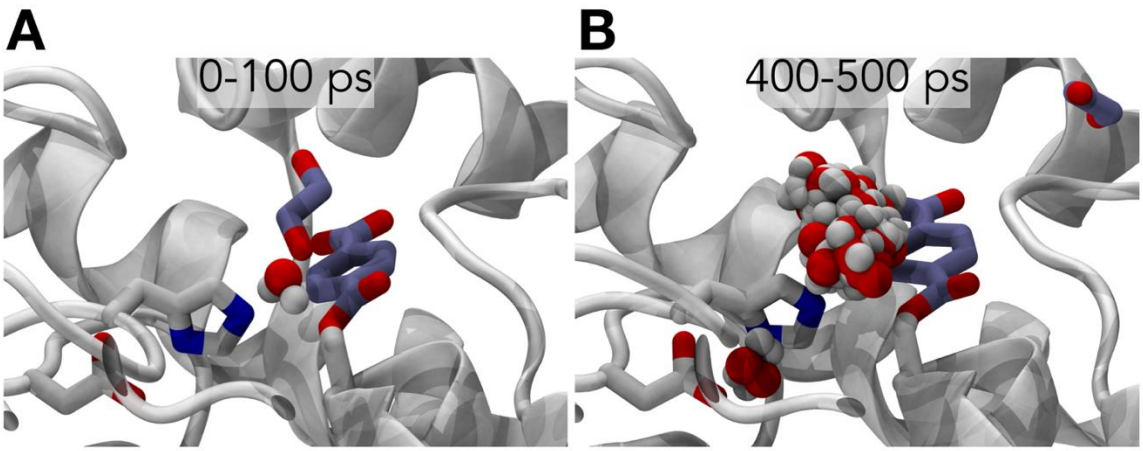


Figure A3.7: Post-acylation simulation of active site and reaction products. Water floods the active site after ethylene glycol (EG) leaves the active site post-acylation, as indicated by molecular dynamics simulations. Three independent MD trajectories with classical forcefield were run of the acyl-enzyme intermediate (AEI) immediately following the first chemical step (acylation). In all three simulations, EG leaves the active site within 4 ns. Results from one such trajectory are shown A) in the first 100 ps after acylation and B) in the time period 400-500 ps after acylation. EG leaves the active site in the intervening time. Water molecules within 3 Å of both the carbonyl carbon of the AEI *and* NE2 atom of His528 are shown every 2 ps. The backbone trace and catalytic residues of MHETase are shown in white cartoon and sticks representation, respectively. Purple sticks show the terephthalic acid moiety of the AEI and EG. Analysis and image were created in VMD.[555]

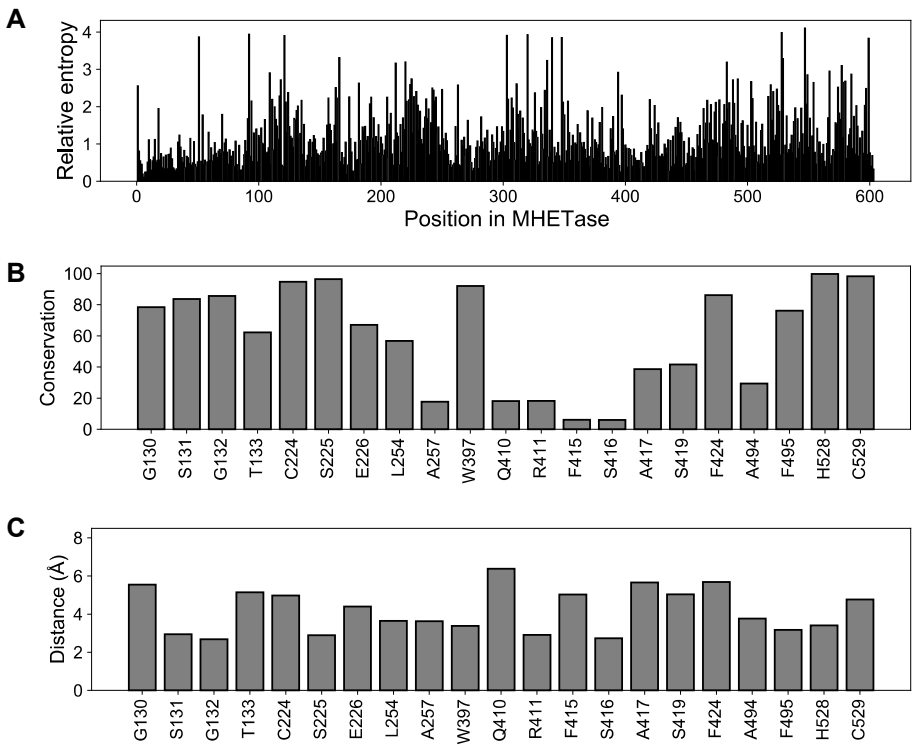


Figure A3.8: Conservation analysis of 6,671 tannase family sequences. A) Conservation scores (relative entropy) of positions in tannase family sequences, plotted against the 603 positions in MHETase. A higher relative entropy implies a greater level of amino acid conservation in the site. B) Conservation scores of active-site residues in MHETase within 6 Å of the MHET substrate, including Gln410 (6.3 Å). Conservation scores are shown as percentiles. Ala257, Gln410, Arg411, Phe415, and Ser416 are the least conserved active-site positions in the active site and are more variable than 81% of all positions in MHETase. C) Closest distance between atoms of MHETase active-site residues and the MHET substrate. The molecular coordinates for MHETase bound with MHET are the same as those in the model from which the molecular simulations were started.

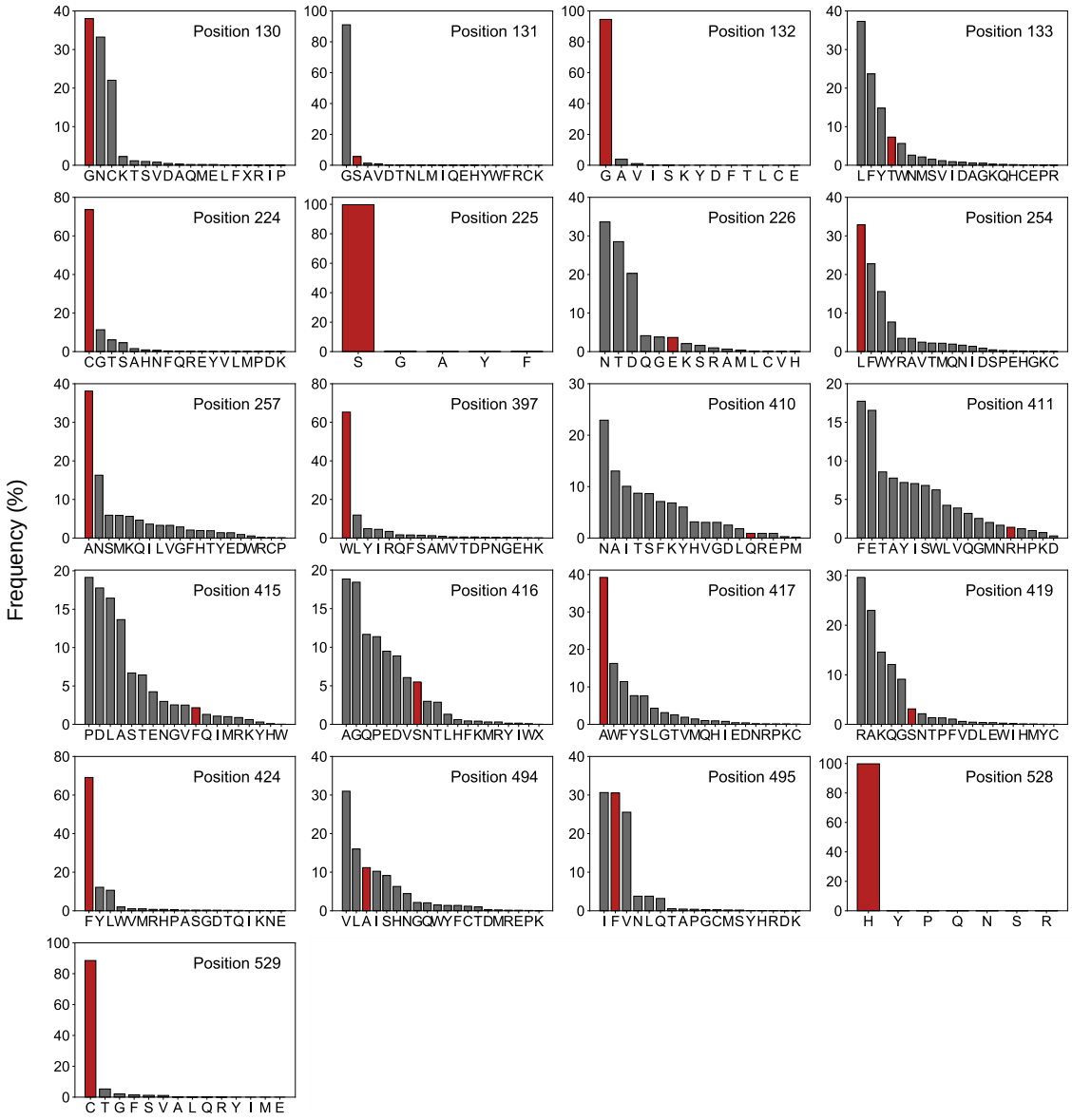


Figure A3.9: Amino acid frequencies of active-site positions in MHETase within 6 Å of the MHET substrate, including Gln410 (6.3 Å). The frequency of amino acids for each position was determined from a MAFFT multiple sequence alignment of 6,671 tannase family sequences. The positions are named using *Is* MHETase numbering, and the red bars indicate the amino acids in *Is* MHETase.

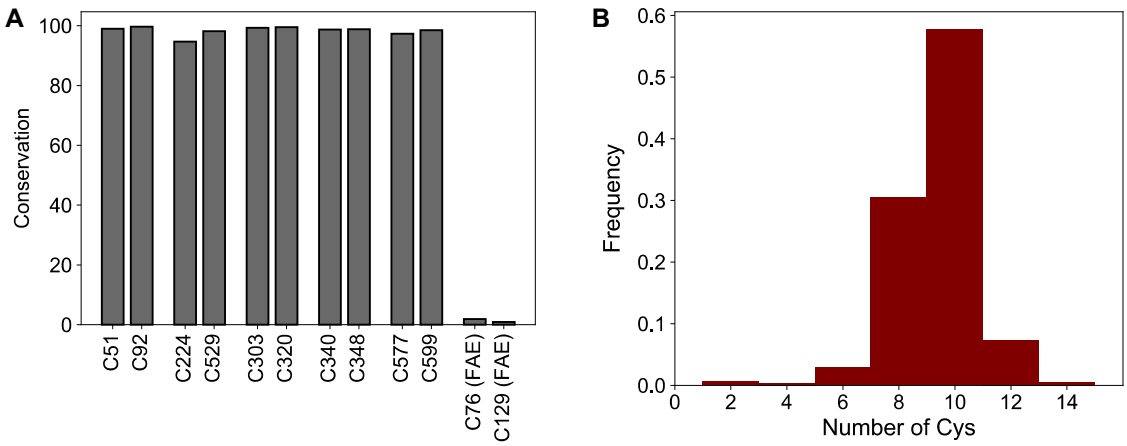


Figure A3.10: Disulfide bond cysteines in 6,671 tannase family sequences. A) Conservation of Cys positions forming five disulfide bonds in MHETase. Conservation scores are shown as percentiles. *Ao* FAEB-1 has a 6th disulfide bond between Cys76 and Cys129 which are very variable positions and are less conserved than 98% of positions in multiple sequence alignment. B) Histogram of Cys occurrence in tannase family sequences showing the rarity of a 6th disulfide bond. Assuming, all Cys form disulfide bonds, less than 8% of tannase family sequences have six disulfide bonds.

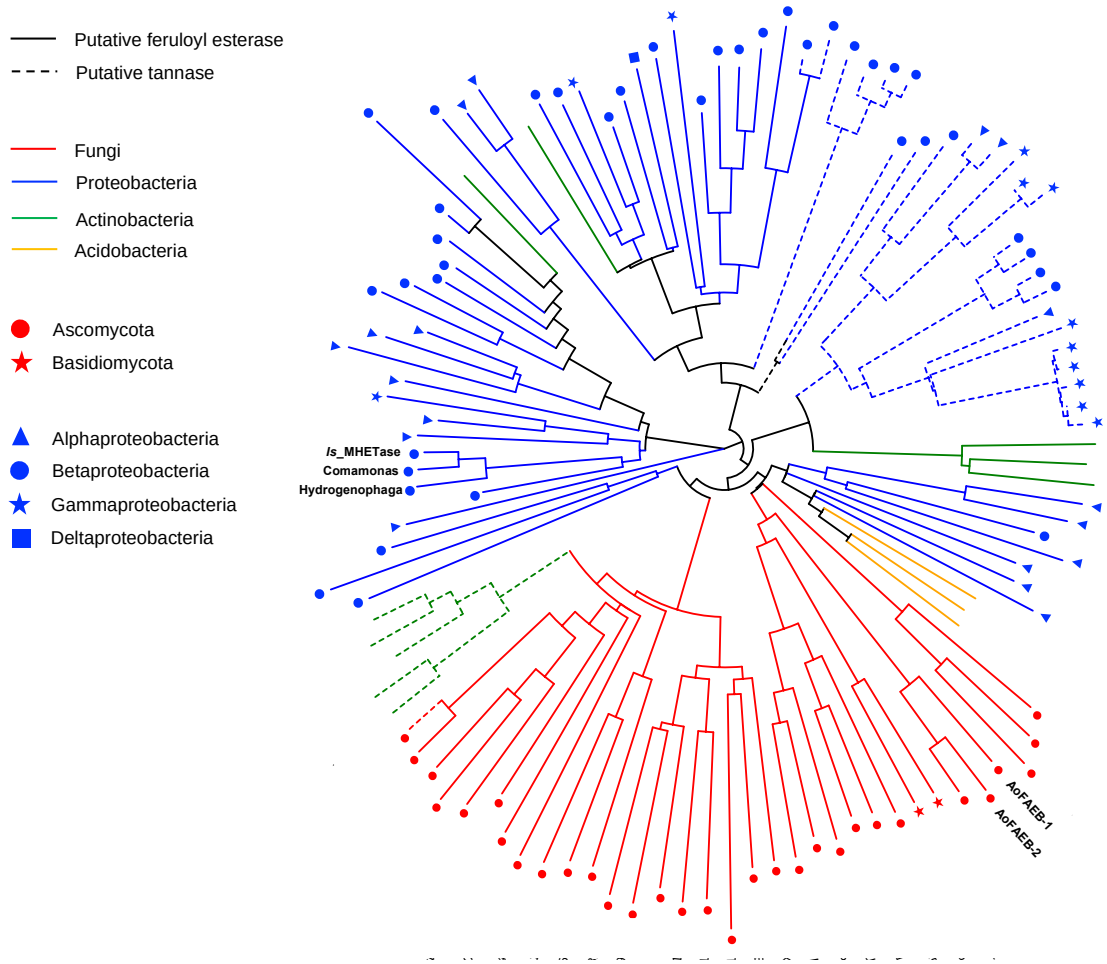


Figure A3.11: Phylogenetic analysis of 120 tannase family sequences with minimum evolution method and 1000 bootstrap replicates. Nodes with bootstrap values between 75% and 100% are indicated with gray circles having sizes that are proportional to the bootstrap values. Multiple sequence alignment was conducted with MAFFT, and the phylogenetic analysis was conducted with MEGA7. *Comamonas* and *Hydrogenophaga* are the close MHETase homolog sequences, with accession codes WP_080747404.1 and WP_083293388.1, respectively. *Ao*FAE-B1 and *Ao*FAE-B2 correspond to the *Aspergillus oryzae* ferulic acid esterases, Q2UP89.1 (PDB 3WMT) and Q2UMX6.1 (PDB 6G21), respectively, which in addition to the recently deposited structures for MHETase (PDB 6QZ1, 6QZ2, 6QZ3, and 6QZ4), the structure from Palm et al. (PDB 6QG9),[85] and

Fusarium oxysporum (PDB 6FAT), are currently the tannase family sequences with solved crystal structures. From the tree, it is clear that FAEs (solid lines) are more phylogenetically similar to MHETase (also shown in solid line) than tannases (dashed lines).

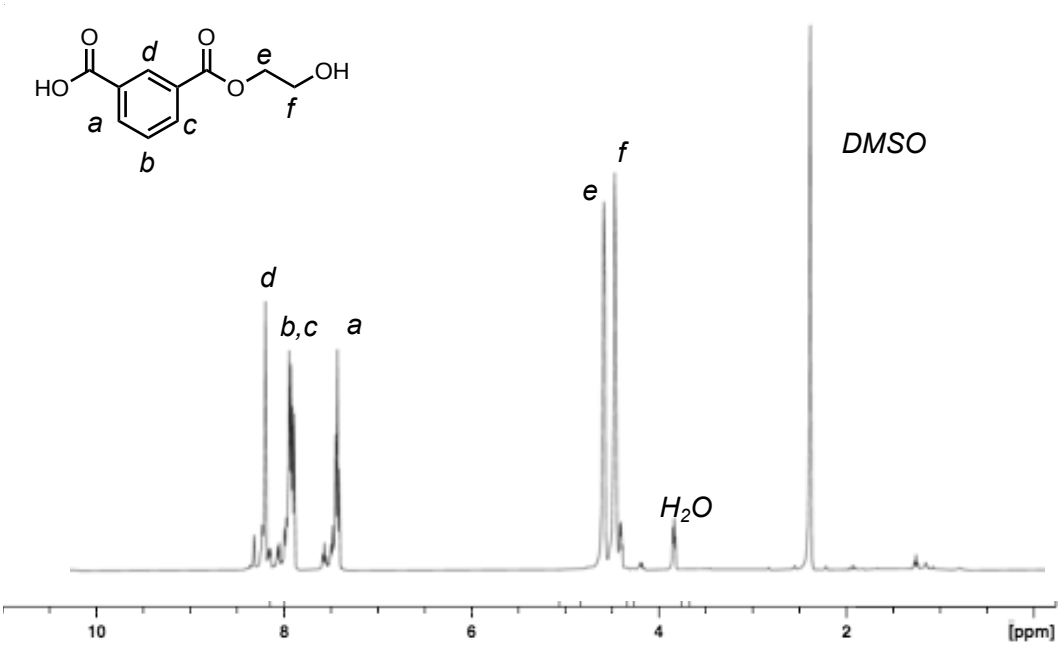


Figure A3.12: Validation by NMR of synthesized mono-(2-hydroxytheyl)-isophthalate. ^{1}H NMR spectrum of MHEI with peak assignments. Integration and peak splitting confirm MHEI was formed and that the product is not BHET.

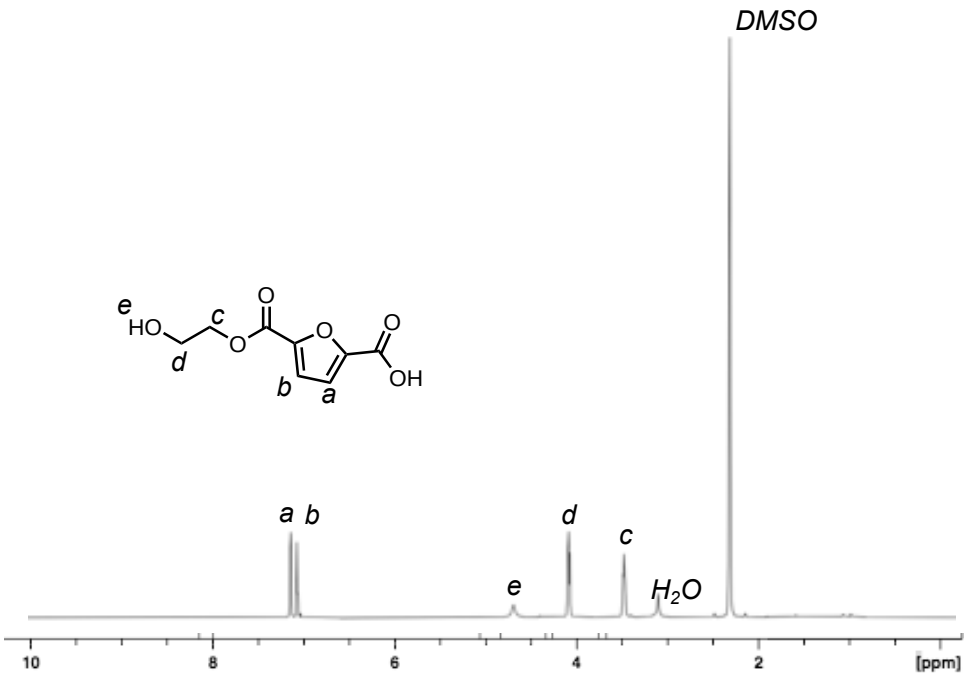


Figure A3.13: Validation by NMR of synthesized mono-(2-hydroxytheyl)-furanoate. ^{1}H NMR spectrum of MHEF with peak assignments. Integration and peak splitting confirm MHEF was formed and that the product is not BHET.

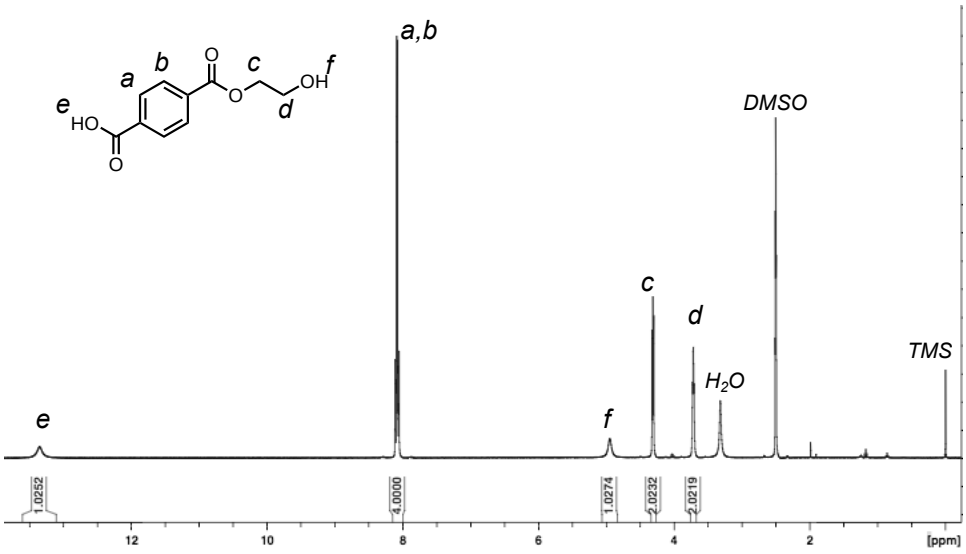


Figure A3.14: Validation by NMR of synthesized mono-(2-hydroxytheyl)-terephthalate. ^{1}H NMR spectrum of MHET with peak assignments. Integration and peak splitting confirm MHET was formed and that the product is not BHET.

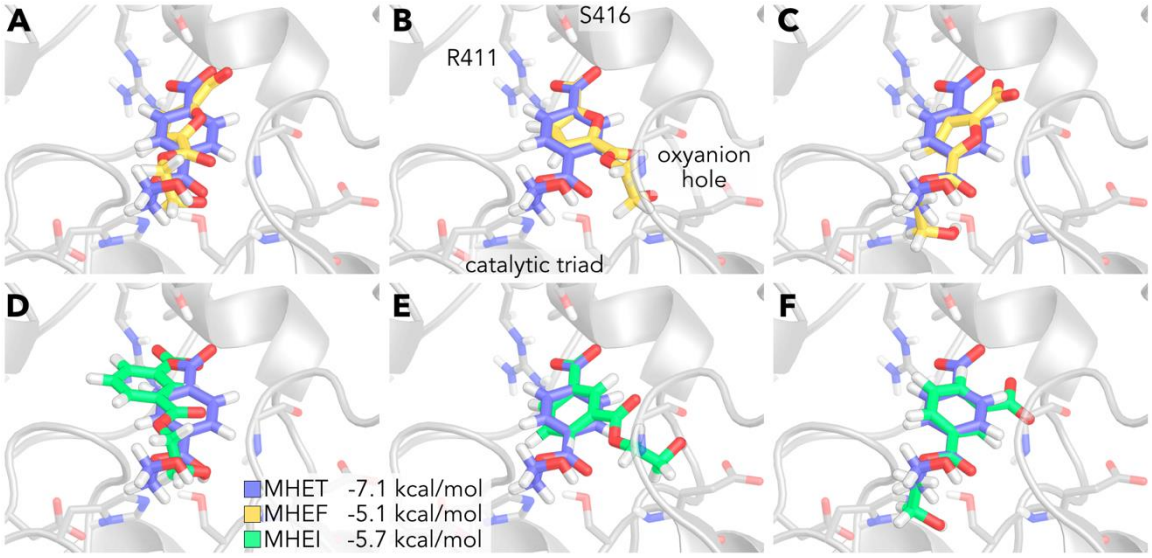


Figure A3.15: Flexible molecular docking studies indicate low energy, catalytically active binding mode for MHET, but not MHEF or MHEI. MHET (purple sticks), MHEF (yellow sticks), MHEI (green sticks) from flexible docking studies with MHETase binding site (grey sticks and ribbons) with ligand visualized in three different ways. A) Enzyme backbones aligned bound with MHET and MHEF. B) Alignment of the carboxylate moiety of MHEF to the carboxylate moiety of MHET, in which case the MHEF carbonyl does not lie in the oxyanion hole (as well as the ester bond being located far from the catalytic residues). C) Alignment of the carbonyl of MHEF to the carbonyl of MHET, in which case the carboxylate is out of range to interact with Arg411. D) Enzyme backbones aligned bound with MHET and MHEI. E) Alignment of the carboxylate moiety of MHEI to the carboxylate moiety of MHET, presenting similar issues as with MHEF. F) Alignment of the carbonyl from MHEI to the carbonyl of MHET. The overlaid binding scores represent the lowest energy binding score for catalytically competent poses (i.e. wherein the catalytic triad was intact).

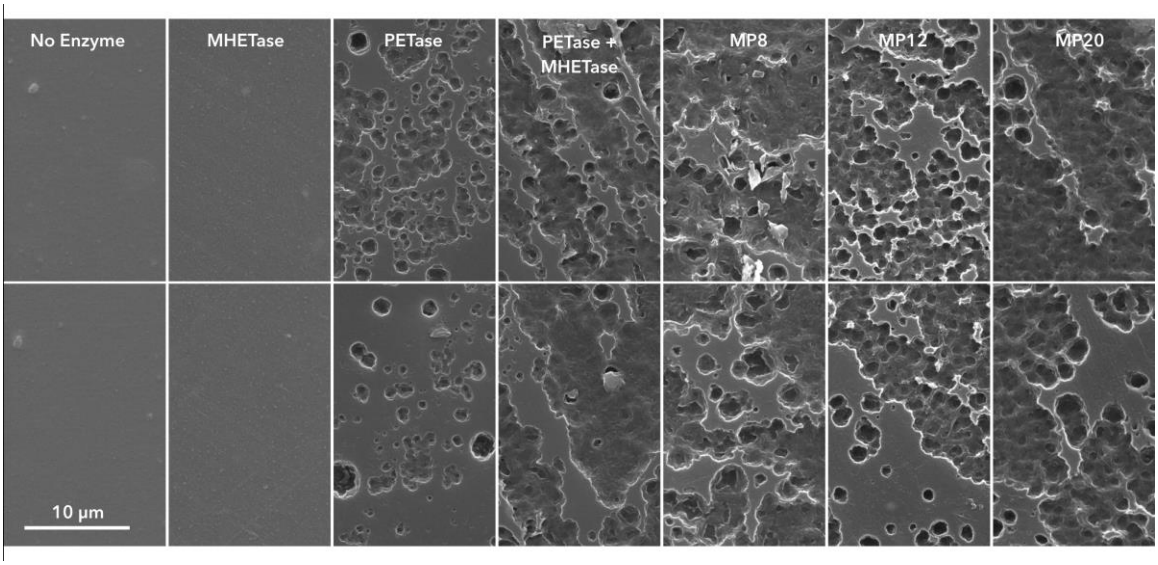


Figure A3.16: SEM of amorphous PET film after 96 h enzyme treatment at 30°C. Digestion conditions represent treatment with no enzyme, treatment with 0.4 mg MHETase/g PET, treatment with 0.4 mg PETase/g PET, simultaneous treatment with 0.4 mg PETase and 0.4 mg MHETase/g PET, and treatment with each chimeric enzyme corresponding to samples presented in Figure A3.4D.

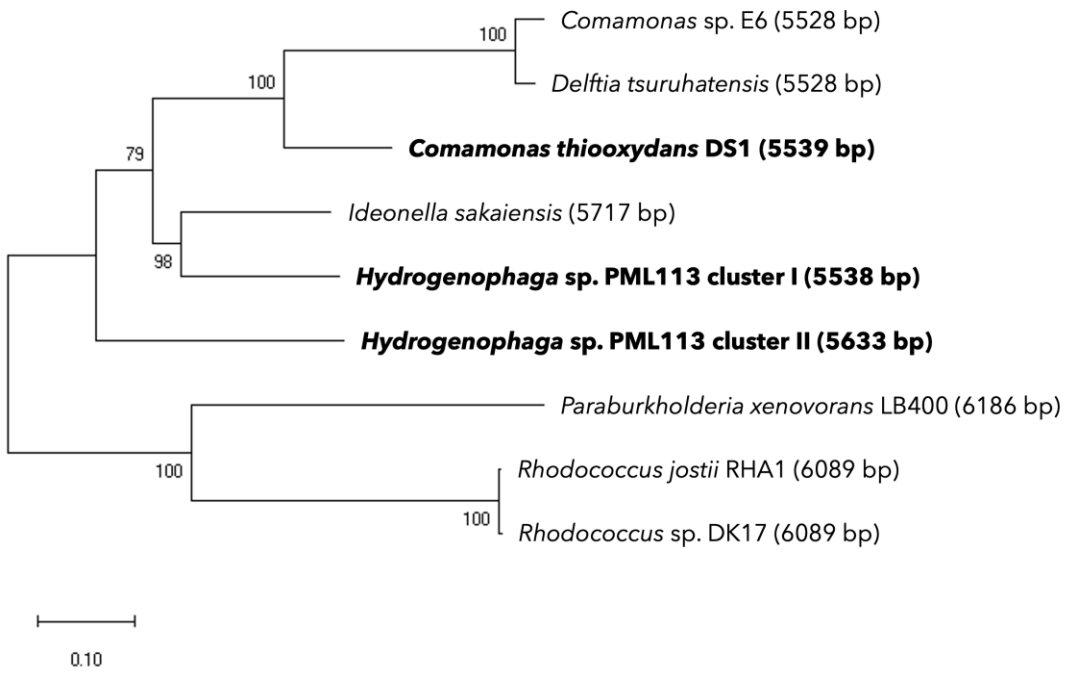


Figure A3.17: Evolutionary analysis by Maximum Likelihood method of known and putative TPA gene clusters. MUSCLE multiple sequence alignment of known and putative TPA gene clusters and prediction of the best evolution model were performed using MEGA X [556]. The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model [172]. The tree with the highest log likelihood (-27899.85) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 0.9655)). The rate variation model allowed for some sites to be evolutionarily invariable ([+*I*], 17.46% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 9 nucleotide sequences. All positions with

less than 95% site coverage were eliminated, i.e., fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position (partial deletion option). There were a total of 4369 positions in the final dataset. Evolutionary analyses were also conducted in MEGA X. Accession numbers for source sequences used are AB238679 for *Comamonas* sp. E6, FOKN01000001 for *Delftia tsuruhatensis*, NZ_AWTM01000090 for *Comamonas thiooxydans* DS1, NZ_BBYR01000104 for *Ideonella sakaiensis*, NZ_MIYM01000023 and NZ_MIYM01000001 for *Hydrogenophaga* sp. PML113 clusters I and II, respectively, CP000271 for *Paraburkholderia xenovorans* LB400, NC_008269 for *Rhodococcus jostii* RHA1, and AY502076 for *Rhodococcus* sp. DK17.

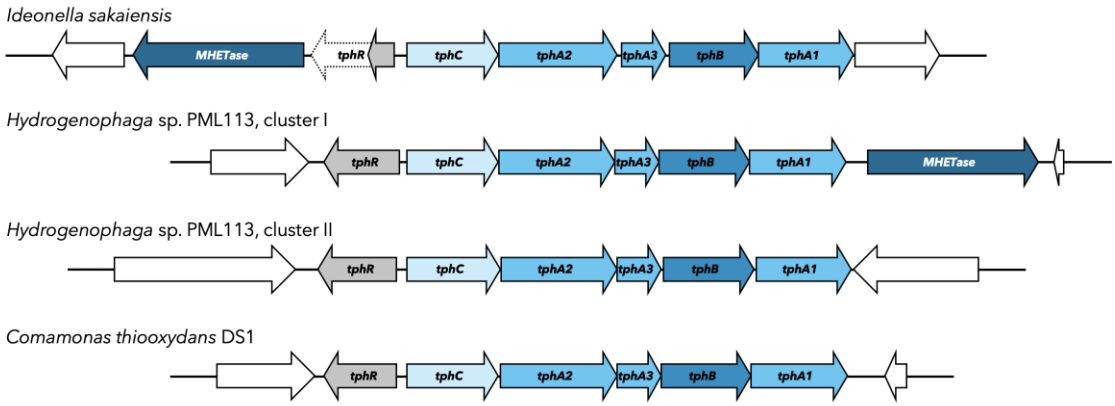


Figure A3.18: Schematic representation of putative TPA catabolic gene clusters in *Hydrogenophaga* sp. PML113 and *Comamonas thiooxydans* DS1, compared to *Ideonella sakaiensis*. A frame-shift in the *I. sakaiensis tphR* coding sequence results in a truncated protein. Searches against the genomes of *C. thiooxydans* strains DF1 and DF2 returned partial sequences due to short contig lengths.

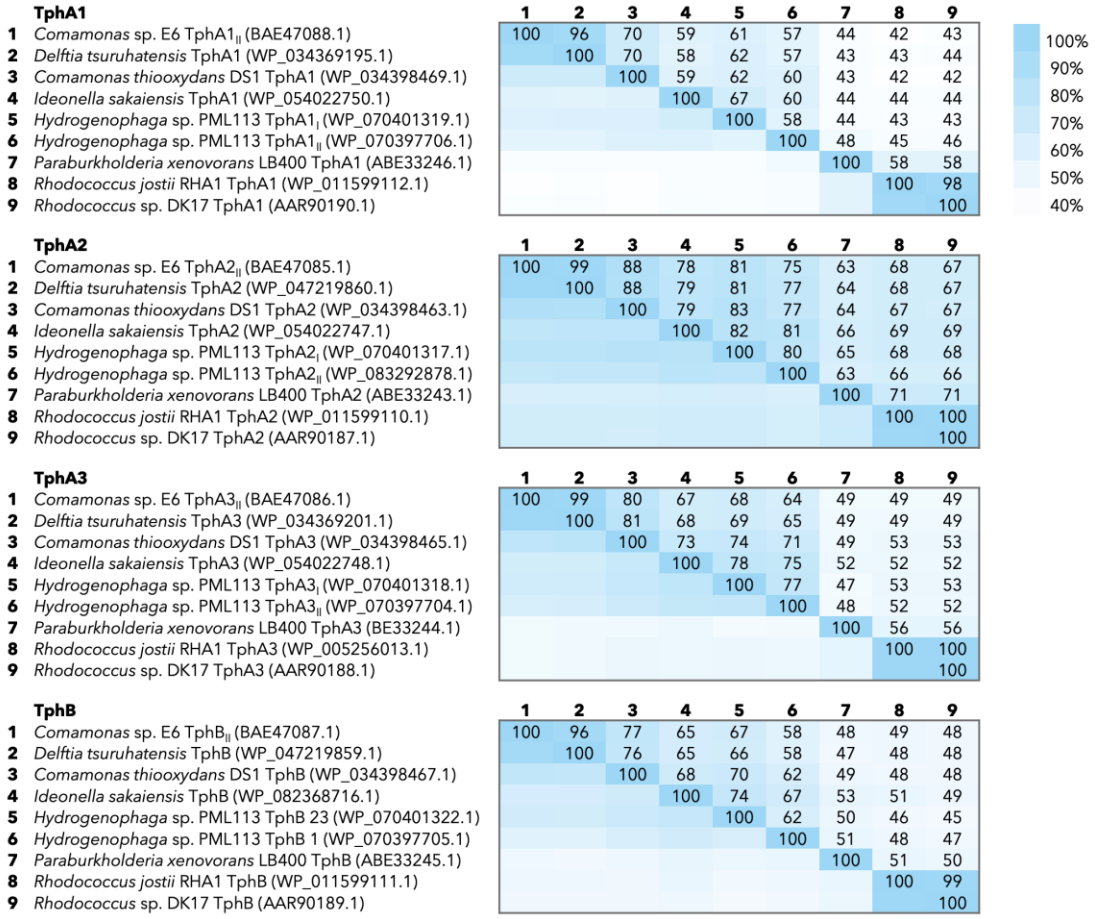


Figure A3.19: Sequence identity matrices for putative TPA catabolic proteins. Identity values obtained from pairwise alignments performed by Clustal Omega.[557]

A3.3 Supplementary Tables

Table A3.1. Crystallographic data, model refinement, and crystallization conditions of *Is*-MHETase.

Data set	SeMet	Native 1	Native 2	Native 3
Space Group	$P22_12_1$	$P22_12_1$	$P2_12_12_1$	$P1$
Wavelength (Å)	0.9795	0.9795	0.9795	0.9795
Resolution Range (Å)	49.30 - 1.60	46.00 - 1.70	46.24 - 1.80	95.29 - 1.90
Unique reflections	84000	70855	122377	506864
Completeness (%)[a]	99.8 (98.9)	99.6 (98.0)	98.2 (96.9)	93.7 (92.4)
Anomalous Completeness (%)[a]	99.8 (98.5)			
*R*merge[b]	0.063 (0.511)	0.056 (0.584)	0.067 (0.550)	0.057 (0.254)
CC(1/2)[c]	0.999 (0.902)	0.999 (0.900)	0.999 (0.914)	
Multiplicity[d]	12.6 (11.0)	6.4 (6.5)	6.7 (6.6)	1.7 (1.7)
Anomalous Multiplicity[d]	6.6 (5.6)			
I/σI[a]	20.2 (4.6)	15.4 (3.1)	14.7 (2.9)	7.1 (2.6)
	a = 77.37 Å,	a = 77.20 Å,	a = 90.21 Å,	a = 110.49 Å,
	b = 89.02 Å,	b = 89.88 Å,	b = 92.80 Å,	b = 135.63 Å,
	c = 91.64 Å	c = 92.00 Å	c = 159.99 Å	c = 138.15 Å,
				α = 83.09°,
				β = 67.91°,
				□ = 67.57°
Model Refinement				
Resolution Range (Å)	45.82 - 1.60	46.00 – 1.70	46.24 - 1.80	46.46 – 1.90
No. of residues:	A: 40-55, 62-603	A: 40-55, 61-603	A: 36-603, B: 36-603	A: 42-603 , B:43-603, C: 43-603, D: 43-603, E: 43-603, F: 43-603, G: 41-603, H: 42-603, I: 43-603, J: 43-603
No. of water, ligands	748, 1 Ca, 1 benzoic acid	552, 1 Ca, 1 benzoic acid	1407, 2 Ca	6125, 10 Ca
Rwork/Rfree (%)[e]	16.20 (17.80)	16.45 (19.18)	18.24 (20.51)	18.54 (20.54)
B average[f]	25.8	32.5	30.9	28.8
Geometry bond, angles[g]	0.003, 0.613	0.008, 0.900	0.005, 0.715	0.003, 0.576
Ramachandran[h]	97.47, 0.0	97.48, 0.2	97.0, 0.0	97.07, 0.02
Molprobity Clash Score	1.61	1.85	3.70	4.15
Beamline	I03	I03	I03	I03
PDB ID[i]	6QZ3	6QZ1	6QZ4	6QZ2

[a] Signal to noise ratio of intensities, highest resolution bin in brackets. [b] R_m : $\sum h \sum i |I/(h,i) - I(h)| / \sum h \sum i / I(h,i)$ where $I(h,i)$ are symmetry-related intensities and $I(h)$ is the mean intensity of the reflection with unique index h . [c] $CC_{1/2}$ is the correlation coefficient of the mean intensities between two random half-datasets. [d] Multiplicity for unique reflections. [e] 5% of

reflections were randomly selected for determination of the free R factor, prior to any refinement. [f] Temperature factors averaged for all atoms. [g] RMS deviations from ideal geometry for bond lengths and restraint angles (9). [h] Percentage of residues in the 'most favoured region' of the Ramachandran plot and percentage of outliers (MOLPROBITY).[558] [i]Protein Data Bank identifiers for coordinates.

Crystallography conditions: SeMet; 0.1 M sodium cacodylate (pH 6.5), 9% PEG 8000. Native 1; 0.1 M sodium acetate (pH 5.5), 24% PEG 5000 MME. Native 2; ammonium acetate (pH 4.5), 22.5% PEG 10000. Native 3; 0.1 M sodium citrate (pH 5.5), 1.0 M ammonium phosphate monobasic. All conditions used 20% glycerol as a cryoprotectant.

Table A3.2. Tannase family sequences used in phylogenetic analysis.

	Accession	Annotation	Organism	Taxon
1	A0A0K8P8E7.1	mono(2-hydroxyethyl) terephthalate hydrolase	*Ideonella sakaiensis*	Betaproteobacteria
2	WP_080747404.1	tannase/feruloyl esterase family alpha/beta	*Comamonas thiooxydans*	Betaproteobacteria
3	WP_083293388.1	tannase/feruloyl esterase family alpha/beta	*Hydrogenophaga sp.*	Betaproteobacteria
4	Q2UP89.1	probable feruloyl esterase b-1	*Aspergillus oryzae*	Ascomycota
5	Q2UMX6.1	Probable feruloyl esterase B-2	*Aspergillus oryzae*	Ascomycota
6	KQO20166.1	feruloyl esterase	*Acidovorax sp.*	Betaproteobacteria
7	SFM74645.1	feruloyl esterase	*Bradyrhizobium sp.*	Alphaproteobacteria
8	EGC99108.1	feruloyl esterase	*Burkholderia sp.*	Betaproteobacteria
9	RLJ38044.1	feruloyl esterase	*Acidovorax sp.*	Betaproteobacteria
10	RKR69440.1	feruloyl esterase	*Acidovorax sp.*	Betaproteobacteria
11	SOD27033.1	feruloyl esterase	*Variovorax sp.*	Betaproteobacteria
12	REF22346.1	feruloyl esterase	*Microbacterium trichothecenolyticum*	Actinobacteria
13	RAR84815.1	feruloyl esterase	*Acidovorax anthurii*	Betaproteobacteria
14	ALV26718.1	feruloyl esterase	*Pannonibacter phragmitetus*	Alphaproteobacteria
15	ODT65815.1	feruloyl esterase	*Pelagibacterium sp.*	Alphaproteobacteria
16	KMO18581.1	feruloyl esterase	*Methylobacterium platani*	Alphaproteobacteria
17	SYX90233.1	putative feruloyl esterase b-1	*Pseudomonas reidholzensis*	Gammaproteobacteria
18	SFO05094.1	feruloyl esterase	*Formivibrio citricus*	Betaproteobacteria
19	SFV19457.1	feruloyl esterase	*Bradyrhizobium arachidis*	Alphaproteobacteria
20	OYX09584.1	feruloyl esterase	*Rhizobiales bacterium*	Alphaproteobacteria
21	OLB33458.1	feruloyl esterase	*Acidobacteria bacterium*	unclassified Acidobacteria
22	OLD21188.1	feruloyl esterase	*Acidobacteria bacterium*	unclassified Acidobacteria
23	SEB13840.1	feruloyl esterase	*Variovorax sp.*	Betaproteobacteria
24	SDU16980.1	feruloyl esterase	*Amycolatopsis keratiniphila*	Actinobacteria
25	SEM06276.1	feruloyl esterase	*Variovorax sp.*	Betaproteobacteria
26	RKT74992.1	feruloyl esterase	*Saccharothrix variisporea*	Actinobacteria
27	PJJ32606.1	feruloyl esterase	*Afipia broomeae*	Alphaproteobacteria
28	CCE09743.1	putative feruloyl esterase	*Bradyrhizobium sp.*	Alphaproteobacteria
29	KVV28346.1	feruloyl esterase	*Burkholderia multivorans*	Betaproteobacteria
30	KXU82530.1	feruloyl esterase	*Paraburkholderia monticola*	Betaproteobacteria
31	OYX06871.1	feruloyl esterase	*Sphingomonadales bacterium*	Alphaproteobacteria
32	ACC69322.1	feruloyl esterase	*Paraburkholderia phymatum*	Betaproteobacteria
33	PVX62318.1	feruloyl esterase	*Sphingomonas sp.*	Alphaproteobacteria
34	ETF04378.1	feruloyl esterase	*Advenella kashmirensis*	Betaproteobacteria
35	KVR34266.1	feruloyl esterase	*Burkholderia ubonensis*	Betaproteobacteria
36	SOB89932.1	feruloyl esterase	*Alcanivorax xenomutans*	Gammaproteobacteria
37	KPV20554.1	feruloyl esterase	*Variovorax paradoxus*	Betaproteobacteria

38	OFW45315.1	feruloyl esterase	*Acidobacteria bacterium*	unclassified Acidobacteria
39	RKD63451.1	feruloyl esterase	*Caballeronia udeis*	Betaproteobacteria
40	SEO24475.1	feruloyl esterase	*Bradyrhizobium sp.*	Alphaproteobacteria
41	ODU08862.1	feruloyl esterase	*Rubrivivax sp.*	Betaproteobacteria
42	KOV79372.1	feruloyl esterase	*Nocardia sp.*	Actinobacteria
43	OYZ97564.1	feruloyl esterase	*Novosphingobium sp.*	Alphaproteobacteria
44	ORY14570.1	feruloyl esterase b precursor	*Clohesyomyces aquaticus*	Ascomycota
45	SCK19622.1	feruloyl esterase	*Variovorax sp.*	Betaproteobacteria
46	PYG17545.1	feruloyl esterase	*Novosphingobium sp.*	Alphaproteobacteria
47	SAK87127.1	feruloyl esterase	*Caballeronia fortuita*	Betaproteobacteria
48	RAR95830.1	feruloyl esterase	*Rahnella sp.*	Gammaproteobacteria
49	AQQ72604.1	feruloyl esterase	*Talaromyces piceae*	Ascomycota
50	SEI21695.1	feruloyl esterase	*Paraburkholderia hospita*	Betaproteobacteria
51	XP_002341414.1	feruloyl esterase, putative	*Talaromyces stipitatus*	Ascomycota
52	PZQ62240.1	feruloyl esterase	*Variovorax paradoxus*	Betaproteobacteria
53	XP_016588953.1	feruloyl esterase	*Sporothrix schenckii*	Ascomycota
54	ODU17436.1	feruloyl esterase	*Variovorax sp.*	Betaproteobacteria
55	ENH84593.1	feruloyl esterase b	*Colletotrichum orbiculare*	Ascomycota
56	SDF78426.1	feruloyl esterase	*Lechevalieria fradiae*	Actinobacteria
57	ESZ94426.1	feruloyl esterase b precursor	*Sclerotinia borealis*	Ascomycota
58	CUA69813.1	putative feruloyl esterase b-2	*Rhizoctonia solani*	Basidiomycota
59	ORY10730.1	feruloyl esterase b precursor	*Clohesyomyces aquaticus*	Ascomycota
60	PIG83702.1	feruloyl esterase b precursor	*Aspergillus arachidicola*	Ascomycota
61	KYF55538.1	feruloyl esterase	*Sorangium cellulosum*	delta/epsilon subdivisions
62	XP_002844880.1	feruloyl esterase b	*Microsporum canis*	Ascomycota
63	KFG78502.1	putative feruloyl esterase	*Metarhizium anisopliae*	Ascomycota
64	SDC72403.1	feruloyl esterase	*Cupriavidus sp.*	Betaproteobacteria
65	XP_018178631.1	feruloyl esterase b	*Purpureocillium lilacinum*	Ascomycota
66	XP_009650693.1	feruloyl esterase b	*Verticillium dahliae*	Ascomycota
67	CEL56090.1	putative feruloyl esterase b-1 os=aspergillus	*Rhizoctonia solani*	Basidiomycota
68	OJW26926.1	feruloyl esterase	*Sphingopyxis sp.*	Alphaproteobacteria
69	XP_013423336.1	putative feruloyl esterase	*Aureobasidium namibiae*	Ascomycota
70	SFT75211.1	feruloyl esterase	*Paraburkholderia aspalathi*	Betaproteobacteria
71	XP_025497975.1	feruloyl esterase b precursor	*Aspergillus aculeatinus*	Ascomycota
72	GAT22098.1	feruloyl esterase b precursor	*Aspergillus luchuensis*	Ascomycota
73	SFG45211.1	feruloyl esterase	*Novosphingobium sp.*	Alphaproteobacteria
74	SIN83151.1	feruloyl esterase	*Paraburkholderia phenazinium*	Betaproteobacteria
75	KNG46319.1	feruloyl esterase b precursor	*Stemphylium lycopersici*	Ascomycota
76	OAQ87654.1	feruloyl esterase	*Purpureocillium lilacinum*	Ascomycota
77	PMD29264.1	putative ferulic acid esterase	*Hyaloscypha variabilis*	Ascomycota
78	GCB21914.1	probable feruloyl esterase arb_07085	*Aspergillus awamori*	Ascomycota
79	XP_001932100.1	feruloyl esterase b precursor	*Pyrenophora tritici-repentis*	Ascomycota

80	EMT66649.1	putative feruloyl esterase b-2	*Fusarium oxysporum*	Ascomycota
81	XP_025508882.1	feruloyl esterase	*Aspergillus aculeatinus*	Ascomycota
82	OAQ76531.1	feruloyl esterase	*Purpureocillium lilacinum*	Ascomycota
83	KLU81119.1	feruloyl esterase b	*Magnaporthiopsis poae*	Ascomycota
84	OMP83047.1	feruloyl esterase b	*Diplodia seriata*	Ascomycota
85	PTD08579.1	putative feruloyl esterase	*Fusarium culmorum*	Ascomycota
86	RKT10778.1	feruloyl esterase	*Paraburkholderia sp.*	Betaproteobacteria
87	XP_025430110.1	putative feruloyl esterase b-2	*Aspergillus saccharolyticus*	Ascomycota
88	XP_025492021.1	feruloyl esterase	*Aspergillus uvarum*	Ascomycota
89	ODU15911.1	feruloyl esterase	*Variovorax sp.*	Betaproteobacteria
90	OXC77937.1	tannase precursor	*Caballeronia sordidicola*	Betaproteobacteria
91	KPX81997.1	tannase	*Pseudomonas meliae*	Gammaproteobacteria
92	RMP50230.1	tannase	*Pseudomonas savastanoi*	Gammaproteobacteria
93	RMV15745.1	tannase	*Pseudomonas savastanoi*	Gammaproteobacteria
94	KPX21055.1	tannase	*Pseudomonas amygdali*	Gammaproteobacteria
95	SPD56196.1	tannase	*Cupriavidus taiwanensis*	Betaproteobacteria
96	BAQ47614.1	tannase	*Methylobacterium aquaticum*	Alphaproteobacteria
97	OTP71705.1	tannase precursor	*Caballeronia sordidicola*	Betaproteobacteria
98	AEA83596.1	tannase precursor	*Pseudomonas stutzeri*	Gammaproteobacteria
99	OUI87529.1	tannase	*Acetobacter sp.*	Alphaproteobacteria
100	KQP45511.1	tannase	*Pseudorhodoferax sp.*	Betaproteobacteria
101	AUB50195.1	tannase precursor	*Klebsiella pneumoniae*	Gammaproteobacteria
102	OAJ68072.1	tannase	*Gluconobacter cerinus*	Alphaproteobacteria
103	ERK18637.1	tannase precursor	*Pantoea sp.*	Gammaproteobacteria
104	CDL19993.1	tannase precursor	*Klebsiella pneumoniae*	Gammaproteobacteria
105	OAG73259.1	tannase	*Gluconobacter japonicus*	Alphaproteobacteria
106	KRC31825.1	tannase	*Acidovorax sp.*	Betaproteobacteria
107	BAU88379.1	tannase	*Streptomyces laurentii*	Actinobacteria
108	KMS92636.1	tannase	*Streptomyces regensis*	Actinobacteria
109	KAK47681.1	tannase	*Caballeronia jiangsuensis*	Betaproteobacteria
110	AKN72753.1	tannase	*Streptomyces sp.*	Actinobacteria
111	OLL31673.1	tannase	*Burkholderia sp.*	Betaproteobacteria
112	KDR33867.1	tannase	*Caballeronia zhejiangensis*	Betaproteobacteria
113	KQB59229.1	tannase	*Acidovorax sp.*	Betaproteobacteria
114	XP_003014523.1	tannase, putative	*Trichophyton benhamiae*	Ascomycota
115	KZT17459.1	tannase	*Acidovorax sp.*	Betaproteobacteria
116	KQO35844.1	tannase	*Acidovorax sp.*	Betaproteobacteria
117	KXU83547.1	tannase	*Paraburkholderia monticola*	Betaproteobacteria
118	KOV87180.1	tannase	*Nocardia sp.*	Actinobacteria
119	KJK45095.1	tannase	*Lechevalieria aerocolonigenes*	Actinobacteria
120	KEZ68139.1	tannase	*Pseudomonas amygdali*	Gammaproteobacteria

Table A3.3. Michaelis-Menten kinetic parameters

Enzyme	K_m	V_{max}	K_i	R^2	k_{cat}/ K_m
	(μM)	(μM s^{-1})	(μM)		(μM^{-1} s^{-1})
Is MHETase	23.17±1.65	0.25±0.05	307.30±20.65	0.90	2.17
Is MHETase S131G	184.10±3.50	0.11±0.03	-	0.93	0.06
Comamonas thiooxydans	174.70±4.75	0.20±0.05	78.80±3.04	0.93	0.23
Hydrogenophaga sp. PML113	41.09±3.38	0.01±0.00	221.50±19.01	0.93	0.13

Results of fitting initial reaction velocities of enzymatic turnover of substrate concentrations between 10 μM and 250 μM using Michaelis-Menten models. The model with substrate inhibition is used for *Is* MHETase, *Comamonas thiooxydans*, and *Hydrogenophage* sp. PML113, while the classic Michaelis-Menten model is used for *Is* MHETase S131G. Non-linear regression performed using GraphPad Prism (8.4.1) is reported, along with 95% confidence intervals for each parameter and R^2 value given for fit of the model to the data.

Table S4. Synergistic degradation of amorphous PET film.

PETase Loading	MHETase Loading	TPA		MHET		BHET		Sum Products	
(mg Enzyme /g PET)		Average (mM)	St. Dev.	Average (mM)	St. Dev.	Average (mM)	St. Dev.	Sum (mM)	St. Dev.
0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0	0.01	0.00	0.02	0.01	0.00	0.00	0.02	0.01
0.1	0.1	0.11	0.03	0.00	0.00	0.00	0.00	0.12	0.03
0.1	0.2	0.19	0.02	0.00	0.00	0.00	0.00	0.19	0.02
0.1	0.3	0.27	0.03	0.00	0.00	0.00	0.00	0.27	0.03
0.1	0.4	0.30	0.05	0.00	0.00	0.00	0.00	0.30	0.05
0.1	0.5	0.27	0.05	0.00	0.00	0.00	0.00	0.28	0.05
0.1	0.6	0.22	0.03	0.00	0.00	0.00	0.00	0.23	0.03
0.1	0.8	0.25	0.02	0.00	0.00	0.00	0.00	0.25	0.02
0.1	1	0.27	0.08	0.00	0.00	0.00	0.00	0.28	0.08
0.2	0	0.03	0.01	0.06	0.01	0.00	0.00	0.10	0.02
0.2	0.1	0.44	0.05	0.00	0.00	0.00	0.00	0.44	0.05
0.2	0.2	0.41	0.12	0.00	0.00	0.00	0.00	0.42	0.12
0.2	0.3	0.44	0.10	0.00	0.00	0.00	0.00	0.45	0.10
0.2	0.4	0.49	0.04	0.00	0.00	0.00	0.00	0.49	0.04
0.2	0.5	0.53	0.06	0.00	0.00	0.00	0.00	0.53	0.06
0.2	0.6	0.40	0.08	0.00	0.00	0.00	0.00	0.40	0.08
0.2	0.8	0.29	0.01	0.00	0.00	0.00	0.00	0.30	0.01
0.2	1	0.37	0.17	0.00	0.00	0.00	0.00	0.37	0.17
0.3	0	0.09	0.00	0.15	0.00	0.00	0.00	0.23	0.01
0.3	0.1	0.61	0.07	0.00	0.00	0.00	0.00	0.61	0.07
0.3	0.2	0.76	0.02	0.00	0.00	0.00	0.00	0.76	0.02
0.3	0.3	0.85	0.08	0.00	0.00	0.00	0.00	0.85	0.08
0.3	0.4	0.80	0.15	0.00	0.00	0.00	0.00	0.81	0.15
0.3	0.5	0.89	0.03	0.00	0.00	0.00	0.00	0.89	0.03
0.3	0.6	0.86	0.13	0.00	0.00	0.00	0.00	0.86	0.13
0.3	0.8	0.89	0.07	0.00	0.00	0.00	0.00	0.89	0.07
0.3	1	0.81	0.12	0.00	0.00	0.00	0.00	0.81	0.12
0.4	0	0.14	0.01	0.19	0.01	0.00	0.00	0.33	0.02
0.4	0.1	0.93	0.15	0.00	0.00	0.00	0.00	0.93	0.15
0.4	0.2	1.03	0.15	0.00	0.00	0.00	0.00	1.03	0.15
0.4	0.3	1.06	0.06	0.00	0.00	0.00	0.00	1.06	0.06
0.4	0.4	1.07	0.04	0.00	0.00	0.00	0.00	1.07	0.04
0.4	0.5	1.06	0.02	0.00	0.00	0.00	0.00	1.06	0.02
0.4	0.6	1.08	0.06	0.00	0.00	0.00	0.00	1.08	0.06
0.4	0.8	1.18	0.06	0.00	0.00	0.00	0.00	1.18	0.06
0.4	1	1.13	0.10	0.00	0.00	0.00	0.00	1.13	0.10
0.5	0	0.20	0.01	0.23	0.01	0.00	0.00	0.43	0.03
0.5	0.1	1.10	0.03	0.00	0.00	0.00	0.00	1.10	0.03
0.5	0.2	1.17	0.07	0.00	0.00	0.00	0.00	1.17	0.07
0.5	0.3	1.35	0.13	0.00	0.00	0.00	0.00	1.35	0.13
0.5	0.4	1.30	0.07	0.00	0.00	0.00	0.00	1.30	0.07
0.5	0.5	1.30	0.10	0.00	0.00	0.00	0.00	1.30	0.10
0.5	0.6	1.32	0.04	0.00	0.00	0.00	0.00	1.32	0.04
0.5	0.8	1.39	0.10	0.00	0.00	0.00	0.00	1.39	0.10
0.5	1	1.27	0.22	0.00	0.00	0.00	0.00	1.27	0.22
0.6	0	0.26	0.02	0.29	0.02	0.00	0.00	0.56	0.05
0.6	0.1	1.35	0.09	0.00	0.00	0.00	0.00	1.35	0.09
0.6	0.2	1.45	0.05	0.00	0.00	0.00	0.00	1.45	0.05
0.6	0.3	1.46	0.16	0.00	0.00	0.00	0.00	1.46	0.16
0.6	0.4	1.52	0.11	0.00	0.00	0.00	0.00	1.52	0.11
0.6	0.5	1.52	0.06	0.00	0.00	0.00	0.00	1.52	0.06

0.6	0.6	1.75	0.18	0.00	0.00	0.00	0.00	1.75	0.18
0.6	0.8	1.66	0.20	0.00	0.00	0.00	0.00	1.66	0.20
0.6	1	1.49	0.04	0.00	0.00	0.00	0.00	1.49	0.04
0.7	0	0.29	0.04	0.27	0.03	0.00	0.00	0.57	0.07
0.7	0.1	1.50	0.13	0.00	0.00	0.00	0.00	1.50	0.13
0.7	0.2	1.51	0.15	0.00	0.00	0.00	0.00	1.51	0.15
0.7	0.3	1.53	0.09	0.00	0.00	0.00	0.00	1.53	0.09
0.7	0.4	1.58	0.05	0.00	0.00	0.00	0.00	1.58	0.05
0.7	0.5	1.46	0.08	0.00	0.00	0.00	0.00	1.46	0.08
0.7	0.6	0.87	0.61	0.00	0.00	0.00	0.00	0.87	0.61
0.7	0.8	1.14	0.48	0.00	0.00	0.00	0.00	1.14	0.48
0.7	1	1.72	0.38	0.00	0.00	0.00	0.00	1.72	0.38
0.8	0	0.34	0.02	0.33	0.01	0.00	0.00	0.67	0.03
0.8	0.1	1.52	0.05	0.00	0.00	0.00	0.00	1.52	0.05
0.8	0.2	1.60	0.39	0.00	0.00	0.00	0.00	1.61	0.40
0.8	0.3	1.80	0.02	0.00	0.00	0.00	0.00	1.80	0.02
0.8	0.4	1.80	0.08	0.00	0.00	0.00	0.00	1.80	0.08
0.8	0.5	1.56	0.24	0.00	0.00	0.00	0.00	1.56	0.24
0.8	0.6	1.78	0.09	0.00	0.00	0.00	0.00	1.79	0.09
0.8	0.8	1.68	0.01	0.00	0.00	0.00	0.00	1.68	0.01
0.8	1	1.82	0.13	0.00	0.00	0.00	0.00	1.82	0.13
0.9	0	0.36	0.01	0.33	0.01	0.00	0.00	0.68	0.01
0.9	0.1	1.58	0.30	0.00	0.00	0.00	0.00	1.58	0.30
0.9	0.2	1.80	0.14	0.00	0.00	0.00	0.00	1.80	0.14
0.9	0.3	2.03	0.09	0.00	0.00	0.00	0.00	2.03	0.09
0.9	0.4	1.95	0.09	0.00	0.00	0.00	0.00	1.95	0.10
0.9	0.5	1.85	0.13	0.00	0.00	0.00	0.00	1.85	0.13
0.9	0.6	1.91	0.05	0.00	0.00	0.00	0.00	1.91	0.05
0.9	0.8	1.56	0.85	0.00	0.00	0.00	0.00	1.56	0.85
0.9	1	1.60	0.23	0.00	0.00	0.00	0.00	1.60	0.23
1	0	0.46	0.03	0.38	0.02	0.00	0.00	0.84	0.05
1	0.1	1.81	0.07	0.00	0.00	0.00	0.00	1.81	0.07
1	0.2	1.83	0.25	0.00	0.00	0.00	0.00	1.83	0.25
1	0.3	1.86	0.30	0.00	0.00	0.00	0.00	1.86	0.30
1	0.4	1.96	0.07	0.00	0.00	0.00	0.00	1.96	0.07
1	0.5	1.94	0.07	0.00	0.00	0.00	0.00	1.94	0.07
1	0.6	2.13	0.07	0.00	0.00	0.00	0.00	2.13	0.07
1	0.8	2.11	0.08	0.00	0.00	0.00	0.00	2.11	0.08
1	1	2.23	0.24	0.00	0.00	0.00	0.00	2.23	0.24
1.2	0	0.55	0.02	0.40	0.02	0.00	0.00	0.95	0.04
1.2	0.1	2.16	0.08	0.00	0.00	0.00	0.00	2.16	0.08
1.2	0.2	2.11	0.05	0.00	0.00	0.00	0.00	2.11	0.05
1.2	0.3	2.28	0.10	0.00	0.00	0.00	0.00	2.29	0.10
1.2	0.4	2.25	0.05	0.00	0.00	0.00	0.00	2.25	0.05
1.2	0.5	2.30	0.05	0.00	0.00	0.00	0.00	2.31	0.05
1.2	0.6	2.32	0.13	0.00	0.00	0.00	0.00	2.32	0.13
1.2	0.8	2.18	0.09	0.00	0.00	0.00	0.00	2.18	0.10
1.2	1	2.26	0.11	0.00	0.00	0.00	0.00	2.26	0.11
1.4	0	0.59	0.03	0.39	0.01	0.00	0.00	0.98	0.04
1.4	0.1	2.21	0.15	0.00	0.00	0.00	0.00	2.21	0.15
1.4	0.2	2.37	0.29	0.00	0.00	0.00	0.00	2.37	0.29
1.4	0.3	2.22	0.11	0.00	0.00	0.00	0.00	2.22	0.11
1.4	0.4	2.25	0.22	0.00	0.00	0.00	0.00	2.26	0.22
1.4	0.5	2.57	0.10	0.00	0.00	0.00	0.00	2.57	0.10
1.4	0.6	2.49	0.01	0.00	0.00	0.00	0.00	2.49	0.01
1.4	0.8	2.49	0.04	0.00	0.00	0.00	0.00	2.49	0.04
1.4	1	2.43	0.06	0.00	0.00	0.00	0.00	2.44	0.06
1.6	0	0.74	0.03	0.41	0.01	0.00	0.00	1.15	0.04
1.6	0.1	2.41	0.16	0.00	0.00	0.00	0.00	2.41	0.16
1.6	0.2	2.46	0.04	0.00	0.00	0.00	0.00	2.46	0.04
1.6	0.3	2.49	0.15	0.00	0.00	0.00	0.00	2.49	0.15
1.6	0.4	2.41	0.10	0.00	0.00	0.00	0.00	2.41	0.10
1.6	0.5	2.70	0.30	0.00	0.00	0.00	0.00	2.71	0.31
1.6	0.6	2.63	0.09	0.00	0.00	0.00	0.00	2.63	0.09
1.6	0.8	2.63	0.16	0.00	0.00	0.00	0.00	2.63	0.16
1.6	1	2.58	0.03	0.00	0.00	0.00	0.00	2.58	0.03

1.8	0	0.81	0.05	0.43	0.03	0.00	0.00	1.24	0.08
1.8	0.1	2.56	0.07	0.00	0.00	0.00	0.00	2.56	0.07
1.8	0.2	2.55	0.07	0.00	0.00	0.00	0.00	2.55	0.07
1.8	0.3	2.57	0.21	0.00	0.00	0.00	0.00	2.57	0.21
1.8	0.4	2.71	0.17	0.00	0.00	0.00	0.00	2.71	0.17
1.8	0.5	2.66	0.02	0.00	0.00	0.00	0.00	2.66	0.02
1.8	0.6	2.60	0.05	0.00	0.00	0.00	0.00	2.60	0.05
1.8	0.8	2.54	0.07	0.00	0.00	0.00	0.00	2.54	0.07
1.8	1	2.65	0.20	0.00	0.00	0.00	0.00	2.65	0.20
2	0	0.87	0.04	0.45	0.02	0.00	0.00	1.32	0.06
2	0.1	2.51	0.11	0.00	0.00	0.00	0.00	2.52	0.11
2	0.2	2.57	0.23	0.00	0.00	0.00	0.00	2.57	0.23
2	0.3	2.66	0.08	0.00	0.00	0.00	0.00	2.66	0.08
2	0.4	2.62	0.26	0.00	0.00	0.00	0.00	2.62	0.26
2	0.5	2.73	0.12	0.00	0.00	0.00	0.00	2.73	0.12
2	0.6	2.66	0.09	0.00	0.00	0.00	0.00	2.66	0.09
2	0.8	2.01	0.37	0.00	0.00	0.00	0.00	2.01	0.37
2	1	2.87	0.13	0.00	0.00	0.00	0.00	2.87	0.13

Reported values represent average and standard deviation for PET constituent monomers released during reactions performed in triplicate over 96 h at 30°C.

Table A3.5. Putative protocatechuate-dioxygenases in *Hydrogenophaga* sp. PML113 and *Comamonas thiooxydans*

Organism	Query	Hit Accession Numbers	% Identity	E-value	Bit score
Hydrogenophaga sp. PML113	LigA	WP_070398564.1	67.2	2.34E-49	166
		WP_070400956.1	43.1	8.31E-27	101
	LigB	WP_070398565.1	63.0	1.35E-122	384
		WP_070400957.1	56.6	5.73E-114	359
Comamonas thiooxydans DS1	LigA	KGH27325.1	65.0	1.25E-46	158
		KGH19511.1	61.7	4.68E-44	151
	LigB	KGH23198.1	62.1	1.45E-76	252
Comamonas thiooxydans DF1	LigA	KGH27550.1	65.0	1.23E-46	158
		KGH13041.1	61.7	4.6E-44	151
	LigB	KGH19836.1 (partial sequence)	69.9	3.14E-56	175
		KGH19529.1 (partial sequence)	62.8	2.23E-47	152
Comamonas thiooxydans DF2	LigA	KGH19350.1	65.0	1.24E-46	158
		KGH20562.1	61.7	4.66E-44	151
	LigB	KGH20561.1	62.3	5.18E-119	374
		KGH19470.1 (partial sequence)	61.4	8E-60	204

Putative protocatechuate (PCA)-dioxygenases in *Hydrogenophaga* sp. PML113 and *Comamonas thiooxydans* strains DS1, DF1, and DF2. PCA-2,3-dioxygenase from *Paenibacillus* sp. JJ-1b (PraA, accession number BAH79099.1), PCA-3,4-dioxygenase alpha and beta subunits from *Pseudomonas putida* KT2440 (PcaH and PcaG, accession numbers WP_010955312.1 and WP_009682255.1) and PCA-4,5-dioxygenase alpha and beta subunits from *Sphingobium* sp. SYK-6 (LigA and LigB, accession numbers BAK65924.1 and BAK65925.1) were used as query. Only hits with >100 bit score from tblastn searches against whole-genome sequences are shown.

Table S6. Summary of conditions tested for quenching MHETase enzymatic activity

Quenching solution	Non-enzymatic hydrolysis of MHET (%)		Enzyme activity quenched?	
	No heat treatment	85°C, 10 min	No heat treatment	85°C, 10 min
20% (v/v) DMSO mixed with 80% (v/v) Buffer Q: 100 mM NaCl, 200 mM sodium phosphate, pH 2.5	0	0	No	No
20% (v/v) DMSO, 80 mM NaCl, 160 mM sodium phosphate, pH 2.5	0	0	No	No
6N HCl, 50% DMSO	4.6	39.4	Unknown (high levels of acidolysis)	Unknown (high levels of acidolysis
100% methanol	0.18	0.25	Yes	Yes
95% ethanol	0	0.69	Yes (causes precipitation)	Yes (causes precipitation)
100% DMSO	0	0	No	No
100 nM PMSF in 100% DMSO	0	1.3	No	Inconsistent
10 mM TCEP in H_2O	0	0.14	No	No
6M GuHCl	0	0.62	No	No
100 nM PMSF in 100% isopropanol	0	0.54	Inconsistent	Inconsistent
6M GuHCl, 10 mM TCEP	0	0.37	No	No

Summary of trial experiments performed in triplicate to determine the most satisfactory method for quenching MHETase enzymatic activity. Experiments were performed in reaction buffer (250 µM MHET, 90 mM NaCl, 10 % (v/v) DMSO, 45 mM sodium phosphate, pH 7.5) and quenched by addition of equal volume of the described quenching solution. The selected quenching method, using 100% methanol and 10 min heat treatment at 85°C, is indicated in grey.

References

1. Bradshaw, C. J.; Brook, B. W., Human population reduction is not a quick fix for environmental problems. *Proc Natl Acad Sci U S A* **2014**, 111, 16610-5.

2. Gerland, P.; Raftery, A. E.; Sevcikova, H.; Li, N.; Gu, D.; Spoorenberg, T.; Alkema, L.; Fosdick, B. K.; Chunn, J.; Lalic, N.; Bay, G.; Buettner, T.; Heilig, G. K.; Wilmoth, J., World population stabilization unlikely this century. *Science* **2014**, 346, 234-7.

3. Li, X.; Zhou, Y.; Eom, J.; Yu, S.; Asrar, G. R., Projecting global urban area growth through 2100 based on historical time series data and future Shared Socioeconomic Pathways. *Earths Future* **2019**, 7, 351-362.

4. Jacobson, M. Z., Review of solutions to global warming, air pollution, and energy security. *Energy Environ. Sci.* **2009**, 2, 148-173.

5. Dincer, I.; Acar, C., A review on clean energy solutions for better sustainability. *Int. J. Energy Res.* **2015**, 39, 585-606.

6. Lucia, L. A., Lignocellulosic biomass: A potential feedstock to replace petroleum. *BioResources* **2008**, 3, 981-982.

7. Johnsson, F.; Kjärstad, J.; Rootzén, J., The threat to climate change mitigation posed by the abundance of fossil fuels. *Clim. Policy* **2019**, 19, 258-274.

8. Meinshausen, M.; Meinshausen, N.; Hare, W.; Raper, S. C.; Frieler, K.; Knutti, R.; Frame, D. J.; Allen, M. R., Greenhouse-gas emission targets for limiting global warming to 2 degrees C. *Nature* **2009**, 458, 1158-62.

9. Marshall, A. L.; Alaimo, P., Useful products from complex starting materials: common chemicals from biomass feedstocks. *Chem. Eur. J.* **2010**, 16, 4970-4980.

10. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D., Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **2007**, 315, 804-7.

11. Chen, C.-C.; Dai, L.; Ma, L.; Guo, R.-T., Enzymatic degradation of plant biomass and synthetic polymers. *Nat. Rev. Chem.* **2020**, 1-13.

12. Payne, C. M.; Knott, B. C.; Mayes, H. B.; Hansson, H.; Himmel, M. E.; Sandgren, M.; Ståhlberg, J.; Beckham, G. T., Fungal cellulases. *Chem. Rev.* **2015**, 115, 1308-1448.

13. Pauly, M.; Keegstra, K., Cell-wall carbohydrates and their modification as a resource for biofuels. *Plant J.* **2008**, 54, 559-568.

14. Fernandes, A. N.; Thomas, L. H.; Altaner, C. M.; Callow, P.; Forsyth, V. T.; Apperley, D. C.; Kennedy, C. J.; Jarvis, M. C., Nanostructure of cellulose microfibrils in spruce wood. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, 108, E1195-203.

15. Klemm, D.; Heublein, B.; Fink, H. P.; Bohn, A., Cellulose: fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Ed. Engl.* **2005**, 44, 3358-93.

16. Pinkert, A.; Marsh, K. N.; Pang, S.; Staiger, M. P., Ionic liquids and their interaction with cellulose. *Chem. Rev.* **2009**, 109, 6712-28.

17. Gardner, K.; Blackwell, J., The structure of native cellulose. *Biopolymers* **1974**, 13, 1975-2001.

18. Nishiyama, Y.; Langan, P.; Chanzy, H., Crystal structure and hydrogen-bonding system in cellulose Iβ from synchrotron X-ray and neutron fiber diffraction. *J. Amer. Chem. Soc.* **2002**, 124, 9074-9082.

19. Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P., Crystal structure and hydrogen bonding system in cellulose Iα from synchrotron X-ray and neutron fiber diffraction. *J. Amer. Chem. Soc.* **2003**, 125, 14300-14306.

20. Atalla, R. H.; Vanderhart, D. L., Native cellulose: a composite of two distinct crystalline forms. *Science* **1984**, 223, 283-285.

21. VanderHart, D. L.; Atalla, R., Studies of microstructure in native celluloses using solid-state carbon-13 NMR. *Macromolecules* **1984**, 17, 1465-1472.

22. Chundawat, S. P.; Beckham, G. T.; Himmel, M. E.; Dale, B. E., Deconstruction of lignocellulosic biomass to fuels and chemicals. *Annu. Rev. Chem. Biomol. Eng.* **2011**.

23. Chundawat, S. P.; Bellesia, G.; Uppugundla, N.; da Costa Sousa, L.; Gao, D.; Cheh, A. M.; Agarwal, U. P.; Bianchetti, C. M.; Phillips Jr, G. N.; Langan, P., Restructuring the crystalline cellulose hydrogen bond network enhances its depolymerization rate. *J. Am. Chem. Soc.* **2011**, 133, 11163-11174.

24. Doi, R. H.; Kosugi, A., Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nat. Rev. Microbiol.* **2004**, 2, 541-551.

25. Ståhlberg, J.; Johansson, G.; Pettersson, G., Trichoderma reesei has no true exo-cellulase: all intact and truncated cellulases produce new reducing end groups on cellulose. *Biochim Biophys Acta* **1993**, 1157, 107-113.

26. Kurašin, M.; Väljamäe, P., Processivity of cellobiohydrolases is limited by the substrate. *J. Biol. Chem.* **2011**, 286, 169-177.

27. Vaaje-Kolstad, G.; Westereng, B.; Horn, S. J.; Liu, Z.; Zhai, H.; Sørlie, M.; Eijsink, V. G., An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. *Science* **2010**, 330, 219-222.

28. Singh, A.; Patel, A. K.; Adsul, M.; Mathur, A.; Singhania, R. R., Genetic modification: a tool for enhancing cellulase secretion. *Biofuel Res. J.* **2017**, 4, 600-610.

29. Puranen, T.; Alapuranen, M.; Vehmaanperä, J. Trichoderma enzymes for textile industries. In *Biotechnology and biology of Trichoderma*; Elsevier: 2014, pp 351-362.

30. Serrano-Ruiz, J. C.; Luque, R.; Sepúlveda-Escribano, A., Transformations of biomass-derived platform molecules: from high added-value chemicals to fuels via aqueous-phase processing. *Chem. Soc. Rev.* **2011**, 40, 5266-5281.

31. www.cazy.org/Glycoside-Hydrolases.html Glycoside Hydrolase family classification.

32. Koshland Jr, D., Stereochemistry and the mechanism of enzymatic reactions. *Biol. Rev.* **1953**, 28, 416-436.

33. Jongkees, S. A.; Withers, S. G., Unusual enzymatic glycoside cleavage mechanisms. *Acc. Chem. Res.* **2014**, 47, 226-235.

34. Ilmen, M.; Saloheimo, A.; Onnela, M.-L.; Penttilä, M. E., Regulation of cellulase gene expression in the filamentous fungus Trichoderma reesei. *Appl. Env. Microbiol.* **1997**, 63, 1298-1306.

35. Momeni, M. H.; Payne, C. M.; Hansson, H.; Mikkelsen, N. E.; Svedberg, J.; Engström, Å.; Sandgren, M.; Beckham, G. T.; Ståhlberg, J., Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus *Heterobasidion irregulare*. *J Biol Chem* **2013**, 288, 5861-5872.

36. Borisova, A. S.; Eneyskaya, E. V.; Bobrov, K. S.; Jana, S.; Logachev, A.; Polev, D. E.; Lapidus, A. L.; Ibatullin, F. M.; Saleem, U.; Sandgren, M.; Payne, C. M.; Kulminskaya, A. A.; Ståhlberg, J., Sequencing, biochemical characterization, crystal structure and molecular dynamics of cellobiohydrolase Cel7A from *Geotrichum candidum* 3C. *FEBS J* **2015**, 282, 4515-4537.

37. Beckham, G. T.; Matthews, J. F.; Bomble, Y. J.; Bu, L.; Adney, W. S.; Himmel, M. E.; Nimlos, M. R.; Crowley, M. F., Identification of amino acids responsible for processivity in a Family 1 carbohydrate-binding module from a fungal cellulase. *J Phys Chem B* **2010**, 114, 1447-1453.

38. Srisodsuk, M.; Lehtiö, J.; Linder, M.; Margolles-Clark, E.; Reinikainen, T.; Teeri, T. T., *Trichoderma reesei* cellobiohydrolase I with an endoglucanase cellulose-binding domain: action on bacterial microcrystalline cellulose. *J Biotechnol* **1997**, 57, 49-57.

39. Borisova, A. S.; Eneyskaya, E. V.; Jana, S.; Badino, S. F.; Kari, J.; Amore, A.; Karlsson, M.; Hansson, H.; Sandgren, M.; Himmel, M. E.; Westh, P.; Payne, C. M.; Kulminskaya, A. A.; Stahlberg, J., Correlation of structure, function and protein dynamics in GH7 cellobiohydrolases from *Trichoderma atroviride*, *T. reese*i and *T. harzianum*. *Biotechnol Biofuels* **2018**, 11, 5.

40. Kurašin, M.; Väljamäe, P., Processivity of cellobiohydrolases is limited by the substrate. *J Biol Chem* **2011**, 286, 169-177.

41. Von Ossowski, I.; Ståhlberg, J.; Koivula, A.; Piens, K.; Becker, D.; Boer, H.; Harle, R.; Harris, M.; Divne, C.; Mahdi, S.; Zhao, Y.; Driguez, H.; Claeyssens, M.; Sinnott, M. L.; Teeri, T. T., Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D. *J Mol Biol* **2003**, 333, 817-829.

42. Hobdey, S. E.; Knott, B. C.; Momeni, M. H.; Taylor, L. E.; Borisova, A. S.; Podkaminer, K. K.; VanderWall, T. A.; Himmel, M. E.; Decker, S. R.; Beckham, G. T.; Stahlberg, J., Biochemical and structural characterizations of two Dictyostelium cellobiohydrolases from the Amoebozoa kingdom reveal a high level of conservation between distant phylogenetic trees of life. *J Appl Environ Microbiol* **2016**, 82, 3395-3409.

43. Sørensen, T. H.; Windahl, M. S.; McBrayer, B.; Kari, J.; Olsen, J. P.; Borch, K.; Westh, P., Loop variants of the thermophile *Rasamsonia emersonii* Cel7A with improved activity against cellulose. *Biotechnol Bioeng* **2017**, 114, 53-62.

44. Ogunmolu, F. E.; Jagadeesha, N. B. K.; Kumar, R.; Kumar, P.; Gupta, D.; Yazdani, S. S., Comparative insights into the saccharification potentials of a relatively unexplored but robust Penicillium funiculosum glycoside hydrolase 7 cellobiohydrolase. *Biotechnol. Biofuels* **2017**, 10, 71.

45. Bu, L.; Beckham, G. T.; Shirts, M. R.; Nimlos, M. R.; Adney, W. S.; Himmel, M. E.; Crowley, M. F., Probing carbohydrate product expulsion from a processive cellulase with multiple absolute binding free energy methods. *J Biol Chem* **2011**, 286, 18161-18169.

46. Taylor, L. E.; Knott, B. C.; Baker, J. O.; Alahuhta, P. M.; Hobdey, S. E.; Linger, J. G.; Lunin, V. V.; Amore, A.; Subramanian, V.; Podkaminer, K.; Xu, Q.-S.; VanderWall, T. A.; Schuster, L. A.; Chaudhari, Y. B.; Adney, W. S.; Crowley, M. F.; Himmel, M. E.; Decker, S. R.; Beckham, G. T., Engineering enhanced cellobiohydrolase activity. *Nat Commun* **2018**, 9, 1186.

47. Goedegebuur, F.; Dankmeyer, L.; Gualfetti, P.; Karkehabadi, S.; Hansson, H.; Jana, S.; Huynh, V.; Kelemen, B. R.; Kruithof, P.; Larenas, E. A.; Teunissen, P. J. M.; Stahlberg, J.; Payne, C. M.; Mitchinson, C.; Sandgren, M., Improving the thermal stability of cellobiohydrolase Cel7A from Hypocrea jecorina by directed evolution. *J. Biol. Chem.* **2017**, 292, 17418-17430.

48. Voutilainen, S. P.; Nurmi-Rantala, S.; Penttila, M.; Koivula, A., Engineering chimeric thermostable GH7 cellobiohydrolases in Saccharomyces cerevisiae. *Appl. Microbiol. Biotechnol.* **2014**, 98, 2991-3001.

49. Sarkanen, K. V.; Ludwig, C. H., *Liguins. Occurrence, formation, structure, and reactions*. 1971.

50. Zakzeski, J.; Bruijnincx, P. C.; Jongerius, A. L.; Weckhuysen, B. M., The catalytic valorization of lignin for the production of renewable chemicals. *Chem. Rev.* **2010**, 110, 3552-3599.

51. Beckham, G. T.; Johnson, C. W.; Karp, E. M.; Salvachúa, D.; Vardon, D. R., Opportunities and challenges in biological lignin valorization. *Current opinion in biotechnology* **2016**, 42, 40-53.

52. Gosselink, R.; E, d.; B, G.; A, A., Coordination network for lignin—3. standardisation, production and applications adapted to market requirements. *Ind. Crops Prod.* **2004**, 20, 121-129.

53. Bugg, T. D.; Ahmad, M.; Hardiman, E. M.; Singh, R., The emerging role for bacteria in lignin degradation and bio-product formation. *Curr Opin Biotechnol* **2011**, 22, 394-400.

54. Floudas, D. B., M.; Riley, R.; Barry, K.; Blanchette, R.A.; Henrissat, B.; Martinez, A.T.; Otillar, R.; Spatafora, J.W.; Yadav, J.S.; Aerts, A.; Benoit, I.; Boyd, A.; Carlson, A.; Copeland, A.; Coutinho, P.M.; de Vries, R.P.; Ferreira, P.; Findley, K.; , The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **2012**, 336, 1715-1719.

55. Machovina, M. M.; Mallinson, S. J. B.; Knott, B. C.; Meyers, A. W.; Garcia-Borras, M.; Bu, L.; Gado, J. E.; Oliver, A.; Schmidt, G. P.; Hinchen, D. J.; Crowley, M. F.; Johnson, C. W.; Neidle, E. L.; Payne, C. M.; Houk, K. N.; Beckham, G. T.; McGeehan, J. E.; DuBois, J. L., Enabling microbial syringol conversion through structure-guided protein engineering. *Proc Natl Acad Sci U S A* **2019**, 116, 13970-13976.

56. Fuchs, G.; Boll, M.; Heider, J., Microbial degradation of aromatic compounds - from one strategy to four. *Nat Rev Microbiol* **2011**, 9, 803-16.

57. Salvachúa, D.; Katahira, R.; Cleveland, N. S.; Khanna, P.; Resch, M. G.; Black, B. A.; Purvine, S. O.; Zink, E. M.; Prieto, A.; Martínez, M. J.; Martínez, A. T.; Simmons, B. A.; Gladden, J. M.; Beckham, G. T., Lignin depolymerization by fungal secretomes and a microbial sink. *Green Chemistry* **2016**, 18, 6046-6062.

58. Whetten, R.; Sederoff, R., Lignin biosynthesis. *Plant Cell* **1995**, 7, 1001.

59. Del Río, J. C.; Rencoret, J.; Prinsen, P.; Martínez, A. n. T.; Ralph, J.; Gutiérrez, A. J. J. o. a., Structural characterization of wheat straw lignin as revealed by analytical pyrolysis, 2D-NMR, and reductive cleavage methods. *J. Agric. Food Chem.* **2012**, 60, 5922-5935.

60. del Río, J. C.; Rencoret, J.; Gutiérrez, A.; Kim, H.; Ralph, J., Hydroxystilbenes are monomers in palm fruit endocarp lignins. *Plant Physiol.* **2017**, 174, 2072-2082.

61. del Río, J. C.; Rencoret, J.; Gutiérrez, A.; Kim, H.; Ralph, J., Structural characterization of lignin from maize (Zea mays L.) fibers: evidence for diferuloylputrescine incorporated into the lignin polymer in maize kernels. *J. Agric. Food Chem.* **2018**, 66, 4402-4413.

62. Rencoret, J.; Neiva, D.; Marques, G.; Gutiérrez, A.; Kim, H.; Gominho, J.; Pereira, H.; Ralph, J.; José, C., Hydroxystilbene glucosides are incorporated into Norway spruce bark lignin. *Plant Physiol.* **2019**, 180, 1310-1321.

63. Ralph, J., Hydroxycinnamates in lignification. *Phytochem. Rev.* **2010**, 9, 65-83.

64. Segura, A.; Bünz, P. V.; D'Argenio, D. A.; Ornston, L. N., Genetic Analysis of a Chromosomal Region ContainingvanA and vanB, Genes Required for Conversion of Either Ferulate or Vanillate to Protocatechuate inAcinetobacter. *J. Bacteriol.* **1999**, 181, 3494-3504.

65. Morawski, B.; Segura, A.; Ornston, L. N., Substrate Range and Genetic Analysis ofAcinetobacter Vanillate Demethylase. *J. Bacteriol.* **2000**, 182, 1383-1389.

66. Yoshikata, T.; Suzuki, K.; Kamimura, N.; Namiki, M.; Hishiyama, S.; Araki, T.; Kasai, D.; Otsuka, Y.; Nakamura, M.; Fukuda, M., Three-component O-demethylase system essential for catabolism of a lignin-derived biphenyl compound in Sphingobium sp. strain SYK-6. *Appl. Env. Microbiol.* **2014**, 80, 7142-7153.

67. Masai, E.; Sasaki, M.; Minakawa, Y.; Abe, T.; Sonoki, T.; Miyauchi, K.; Katayama, Y.; Fukuda, M., A novel tetrahydrofolate-dependent O-demethylase gene is essential for growth of Sphingomonas paucimobilis SYK-6 with syringate. *J. Bacteriol.* **2004**, 186, 2757-2765.

68. Abe, T.; Masai, E.; Miyauchi, K.; Katayama, Y.; Fukuda, M., A tetrahydrofolate-dependent O-demethylase, LigM, is crucial for catabolism of vanillate and syringate in Sphingomonas paucimobilis SYK-6. *Journal of Bacteriology* **2005**, 187, 2030-2037.

69. Dardas, A. G., D.; Barrelle, M.; Sauret-Ignazi, G.; Sterjiades, R.; Pelmont, J., The demethylation of guaiacol by a new bacterial cytochrome P-450. *Arch Biochem Biophys* **1985**, 236, 585-592.

70. Mallinson, S. J. B.; Machovina, M. M.; Silveira, R. L.; Garcia-Borras, M.; Gallup, N.; Johnson, C. W.; Allen, M. D.; Skaf, M. S.; Crowley, M. F.; Neidle, E. L.; Houk, K. N.; Beckham, G. T.; DuBois, J. L.; McGeehan, J. E., A promiscuous cytochrome P450 aromatic O-demethylase for lignin bioconversion. *Nat Commun* **2018**, 9, 2487.

71. Mandlekar, N.; Cayla, A.; Rault, F.; Giraud, S.; Salaün, F.; Malucelli, G.; Guan, J.-P., An overview on the use of lignin and its derivatives in fire retardant polymer systems. *Lignin-Trends and Applications* **2018**.

72. Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.; Toyohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K., A bacterium that degrades and assimilates poly (ethylene terephthalate). *Science* **2016**, 351, 1196-1199.

73. Müller, R.-J.; Kleeberg, I.; Deckwer, W.-D., Biodegradation of polyesters containing aromatic constituents. *J. Biotechnol.* **2001**, 86, 87-95.

74. Geyer, R.; Jambeck, J. R.; Law, K. L., Production, use, and fate of all plastics ever made. *Sci. Adv.* **2017**, 3, e1700782.

75. Garcia, J. M.; Robertson, M. L., The future of plastics recycling. *Science* **2017**, 358, 870-872.

76. Rahimi, A.; García, J. M., Chemical recycling of waste plastics for new materials production. *Nat. Rev. Chem.* **2017**, 1, 1-11.

77. Chen, C.-C.; Dai, L.; Ma, L.; Guo, R.-T. J. N. R. C., Enzymatic degradation of plant biomass and synthetic polymers. **2020**, 1-13.

78. Taniguchi, I.; Yoshida, S.; Hiraga, K.; Miyamoto, K.; Kimura, Y.; Oda, K., Biodegradation of PET: Current Status and Application Aspects. *ACS Catal.* **2019**, 9, 4089-4105.

79. Danso, D.; Schmeisser, C.; Chow, J.; Zimmermann, W.; Wei, R.; Leggewie, C.; Li, X.; Hazen, T.; Streit, W. R., New insights into the function and global distribution of polyethylene terephthalate (PET)-degrading bacteria and enzymes in marine and terrestrial metagenomes. *Appl. Environ. Microbiol.* **2018**, 84, e02773-17.

80. Tournier, V.; Topham, C.; Gilles, A.; David, B.; Folgoas, C.; Moya-Leclair, E.; Kamionka, E.; Desrousseaux, M.-L.; Texier, H.; Gavalda, S.; Cot, M.; Guemard, E.; Dalibey, M.; Nomme, J.; Cioci, G.; Barbe, S.; Chateau, M.; Andre, I.; Duquesne, S.; Marty, A., An engineered PET depolymerase to break down and recycle plastic bottles. **2020**, 580, 216-219.

81. Austin, H. P.; Allen, M. D.; Donohoe, B. S.; Rorrer, N. A.; Kearns, F. L.; Silveira, R. L.; Pollard, B. C.; Dominick, G.; Duman, R.; El Omari, K., Characterization and

engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci.* **2018**, 115, E4350-E4357.

82. Joo, S.; Cho, I. J.; Seo, H.; Son, H. F.; Sagong, H.-Y.; Shin, T. J.; Choi, S. Y.; Lee, S. Y.; Kim, K.-J., Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation. *Nature Comm.* **2018**, 9, 382.

83. Liu, B.; He, L.; Wang, L.; Li, T.; Li, C.; Liu, H.; Luo, Y.; Bao, R., Protein crystallography and site-direct mutagenesis analysis of the poly (ethylene terephthalate) hydrolase PETase from Ideonella sakaiensis. *ChemBioChem* **2018**, 19, 1471-1475.

84. Cui, Y.; Chen, Y.; Liu, X.; Dong, S.; Qiao, Y.; Han, J.; Li, C.; Han, X.; Liu, W.; Chen, Q., Computational redesign of PETase for plastic biodegradation by GRAPE strategy. *BioRxiv* **2019**, 787069.

85. Palm, G. J.; Reisky, L.; Böttcher, D.; Müller, H.; Michels, E. A. P.; Walczak, M. C.; Berndt, L.; Weiss, M. S.; Bornscheuer, U. T.; Weber, G., Structure of the plastic-degrading *Ideonella sakaiensis* MHETase bound to a substrate. *Nature Comm.* **2019**, 10, 1717.

86. Sagong, H.-Y.; Seo, H.; Kim, T.; Son, H. F.; Joo, S.; Lee, S. H.; Kim, S.; Woo, J.-S.; Hwang, S. Y.; Kim, K.-J., Decomposition of PET film by MHETase using Exo-PETase function. *ACS Catal.* **2020**, In press.

87. Mjolsness, E.; DeCoste, D., Machine learning for science: state of the art and future prospects. *Science* **2001**, 293, 2051-5.

88. Alpaydin, E., *Introduction to machine learning*. MIT press: 2020.

89. Guzella, T. S.; Caminhas, W. M., A review of machine learning approaches to spam filtering. *Expert Syst. Appl.* **2009**, 36, 10206-10222.

90. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; 2016; pp 770-778.

91. Dong, D.; Wu, H.; He, W.; Yu, D.; Wang, H. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015; 2015; pp 1723-1732.

92. De Bruijne, M., In; Elsevier: 2016.

93. Polson, N. G.; Sokolov, V. O., Deep learning for short-term traffic flow prediction. *Transp. Res. Part C Emerg. Technol.* **2017**, 79, 1-17.

94. Perlich, C.; Dalessandro, B.; Raeder, T.; Stitelman, O.; Provost, F., Machine learning for targeted display advertising: Transfer learning in action. *Mach. Learn.* **2014**, 95, 103-127.

95. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N., Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, 566, 195-204.

96. Larranaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J. A.; Armañanzas, R.; Santafé, G.; Pérez, A., Machine learning in bioinformatics. *Brief. Bioinformatics* **2006**, 7, 86-112.

97. Baldi, P.; Brunak, S.; Bach, F., *Bioinformatics: the machine learning approach*. MIT press: 2001.

98. Tarca, A. L.; Carey, V. J.; Chen, X.-w.; Romero, R.; Drăghici, S., Machine learning and its applications to biology. *PLOS Comput. Biol.* **2007**, 3, e116.

99. Schölkopf, B.; Tsuda, K.; Vert, J.-P., *Kernel methods in computational biology*. MIT press: 2004.

100. Sommer, C.; Gerlich, D. W., Machine learning in cell biology–teaching computers to recognize phenotypes. *J. Cell Sci.* **2013**, 126, 5529-5539.

101. Swan, A. L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J., Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS* **2013**, 17, 595-610.

102. Bengio, Y.; Courville, A.; Vincent, P., Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell* **2013**, 35, 1798-1828.

103. Bengio, Y.; Goodfellow, I.; Courville, A., *Deep learning*. MIT press Massachusetts, USA:: 2017; Vol. 1.

104. Li, H.; Tian, S.; Li, Y.; Fang, Q.; Tan, R.; Pan, Y.; Huang, C.; Xu, Y.; Gao, X., Modern deep learning in bioinformatics. *J. Mol. Cell Biol.* **2020**.

105. Ng, A., Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. *Retrieved online at https://www.mlyearning.org* **2019**.

106. Chicco, D., Ten quick tips for machine learning in computational biology. *BioData Min* **2017**, 10, 35.

107. Kim, J.-H., Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal* **2009**, 53, 3735-3745.

108. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Matthieu, P.; Duchesnay, E., Scikit-learn: Machine learning in Python. *J Mach Learn Res* **2011**, 12, 2825-2830.

109. Chicco, D.; Jurman, G., The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, 21, 6.

110. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R., Small-sample precision of ROC-related estimates. *Bioinformatics* **2010**, 26, 822-830.

111. Lobo, J. M.; Jiménez-Valverde, A.; Real, R., AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **2008**, 17, 145-151.

112. Hand, D. J., Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach. Learn.* **2009**, 77, 103-123.

113. He, H.; Garcia, E. A., Learning from imbalanced data. *IEEE T. Knowl. Data. En.* **2009**, 21, 1263-1284.

114. Krawczyk, B., Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **2016**, 5, 221-232.

115. Branco, P.; Torgo, L.; Ribeiro, R. P., A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* **2016**, 49, 1-50.

116. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P., SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, 16, 321-357.

117. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N. V., SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, 61, 863-905.

118. Branco, P.; Torgo, L.; Ribeiro, R. P., Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing* **2019**, 343, 76-99.

119. Branco, P.; Ribeiro, R. P.; Torgo, L., UBL: an R package for utility-based learning. *arXiv preprint arXiv:1604.08079* **2016**.

120. Branco, P.; Torgo, L.; Ribeiro, R. P. SMOGN: a Pre-processing Approach for Imbalanced Regression. In First International Workshop on Learning with Imbalanced Domains: Theory and Applications, 2017; 2017; pp 36-50.

121. Branco, P.; Torgo, L.; Ribeiro, R. P. Rebagg: Resampled bagging for imbalanced regression. In Second International Workshop on Learning with Imbalanced Domains: Theory and Applications, 2018; 2018; pp 67-81.

122. Torgo, L.; Branco, P.; Ribeiro, R. P.; Pfahringer, B., Resampling strategies for regression. *Expert Syst.* **2015**, 32, 465-476.

123. Kleinbaum, D. G.; Kupper, L. L.; Chambless, L. E., Logistic regression analysis of epidemiologic data: theory and practice. *Commun. Stat. Theory Methods* **1982**, 11, 485-547.

124. Fix, E., *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine: 1951.

125. Vapnik, V. N., The nature of statistical learning theory. **1995**.

126. Serra, A.; Galdi, P.; Tagliaferri, R., Machine learning for bioinformatics and neuroimaging. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, 8, e1248.

127. Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A., *Classification and regression trees*. CRC press: 1984.

128. Quinlan, J. R., *C4. 5: programs for machine learning*. Elsevier: 2014.

129. Quinlan, J. R., Induction of decision trees. *Mach. Learn.* **1986**, 1, 81-106.

130. Breiman, L., Bagging predictors. *Mach. Learn.* **1996**, 24, 123-140.

131. Breiman, L., Random forests. *Mach Learn* **2001**, 45, 5-32.

132. Qi, Y. Random forest for bioinformatics. In *Ensemble machine learning*; Springer: 2012, pp 307-323.

133. Kandaswamy, K. K.; Chou, K. C.; Martinetz, T.; Moller, S.; Suganthan, P. N.; Sridharan, S.; Pugalenthi, G., AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J Theor Biol* **2011**, 270, 56-62.

134. Han, P.; Zhang, X.; Norton, R. S.; Feng, Z.-P., Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinf* **2009**, 10, 8.

135. Chen, X.; Ishwaran, H., Random forests for genomic data analysis. *Genomics* **2012**, 99, 323-329.

136. Archer, K. J.; Kimes, R. V., Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal* **2008**, 52, 2249-2260.

137. Mount, D., Bioinformatics: sequence and genome analysis. 2004. *Cold Spring Harbor Laboratory Press*.

138. Chatzou, M.; Magis, C.; Chang, J.-M.; Kemena, C.; Bussotti, G.; Erb, I.; Notredame, C., Multiple sequence alignment modeling: methods and applications. *Brief. Bioinformatics* **2016**, 17, 1009-1023.

139. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Res.* **1994**, 22, 4673-4680.

140. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, 7, 539.

141. Notredame, C.; Higgins, D. G.; Heringa, J., T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, 302, 205-217.

142. Do, C. B.; Mahabhashyam, M. S.; Brudno, M.; Batzoglou, S., ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **2005**, 15, 330-340.

143. Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Res.* **2004**, 32, 1792-1797.

144. Katoh, K.; Standley, D. M., MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **2013**, 30, 772-780.

145. Capra, J. A.; Singh, M., Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007**, 23, 1875-1882.

146. Petrova, N. V.; Wu, C. H., Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinf.* **2006**, 7, 312.

147. Liang, S.; Zhang, C.; Liu, S.; Zhou, Y., Protein binding site prediction using an empirical scoring function. *Nucleic Acid Res.* **2006**, 34, 3698-3707.

148. Magliery, T. J.; Regan, L., Sequence variation in ligand binding sites in proteins. *BMC Bioinf.* **2005**, 6, 1-11.

149. Caffrey, D. R.; Somaroo, S.; Hughes, J. D.; Mintseris, J.; Huang, E. S., Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **2004**, 13, 190-202.

150. Guharoy, M.; Chakrabarti, P., Conservation and relative importance of residues across protein-protein interfaces. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, 102, 15447-15452.

151. Mintseris, J.; Weng, Z., Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, 102, 10930-10935.

152. Hannenhalli, S. S.; Russell, R. B., Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* **2000**, 303, 61-76.

153. Kalinina, O. V.; Mironov, A. A.; Gelfand, M. S.; Rakhmaninova, A. B., Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.* **2004**, 13, 443-456.

154. Lichtarge, O.; Bourne, H. R.; Cohen, F. E., An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **1996**, 257, 342-358.

155. Karlin, S.; Brocchieri, L., Evolutionary conservation of RecA genes in relation to protein structure and function. *J. Bacteriol.* **1996**, 178, 1881-1894.

156. Schueler-Furman, O.; Baker, D., Conserved residue clustering and protein structure prediction. *Proteins* **2003**, 52, 225-235.

157. Valdar, W. S.; Thornton, J. M., Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **2001**, 313, 399-416.

158. Johansson, F.; Toh, H., A comparative study of conservation and variation scores. *BMC Bioinf.* **2010**, 11, 388.

159. Lockless, S. W.; Ranganathan, R., Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **1999**, 286, 295-299.

160. Lin, J., Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, 37, 145-151.

161. Mayrose, I.; Graur, D.; Ben-Tal, N.; Pupko, T., Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *J. Mol. Biol.* **2004**, 21, 1781-1791.

162. Sander, C.; Schneider, R., Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, 9, 56-68.

163. Wu, T. T.; Kabat, E. A., An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.* **1970**, 132, 211-250.

164. Landgraf, R.; Fischer, D.; Eisenberg, D., Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* **1999**, 12, 943-951.

165. Mihalek, I.; Reš, I.; Lichtarge, O., A family of evolution–entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.* **2004**, 336, 1265-1282.

166. Ashkenazy, H.; Erez, E.; Martz, E.; Pupko, T.; Ben-Tal, N., ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acid Res.* **2010**, 38, W529-W533.

167. Kreft, L.; Turan, D.; Hulstaert, N.; Botzki, A.; Martens, L.; Vandermarliere, E., Scop3D: Online Visualization of Mutation Rates on Protein Structure. *J. Proteome Res.* **2018**, 18, 765-769.

168. Płuciennik, A.; Stolarczyk, M.; Bzówka, M.; Raczyńska, A.; Magdziarz, T.; Góra, A., BALCONY: an R package for MSA and functional compartments of protein variability analysis. *BMC Bioinf.* **2018**, 19, 1-8.

169. Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, 25, 1422-1423.

170. Mayr, E., The role of systematics in biology. *Science* **1968**, 159, 595-9.

171. Darwin, C., On the origin of species London. *John Murray, London* **1859**, 62.

172. Nei, M.; Kumar, S., *Molecular evolution and phylogenetics*. Oxford university press: 2000.

173. Cavalli-Sforza, L. L.; Edwards, A. W., Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **1967**, 19, 233.

174. Felsenstein, J., *Inferring phylogenies*. Sinauer associates Sunderland, MA: 2004; Vol. 2.

175. Sober, E., The contest between parsimony and likelihood. *Syst. Biol.* **2004**, 53, 644-653.

176. Lemey, P.; Salemi, M.; Vandamme, A.-M., *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press: 2009.

177. Uzzell, T.; Corbin, K. W., Fitting discrete probability distributions to evolutionary events. *Science* **1971**, 172, 1089-1096.

178. Grishin, N. V., Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* **1995**, 41, 675-679.

179. Dayhoff, M. O., A model of evolutionary change in proteins. *Atlas of protein sequence and structure* **1972**, 5, 89-99.

180. Jones, D. T.; Taylor, W. R.; Thornton, J. M., The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **1992**, 8, 275-282.

181. Vinh, L. S.; von Haeseler, A., Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC Bioinf.* **2005**, 6, 92.

182. Rzhetsky, A.; Nei, M., A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **1992**.

183. Felsenstein, J., Confidence intervals on phylogenetics: an approach using bootstrap. =. *Evolution* **1985**, 39, 783–791.

184. Efron, B.; Halloran, E.; Holmes, S., Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, 93, 13429-13429.

185. Payne, C. M.; Knott, B. C.; Mayes, H. B.; Hansson, H.; Himmel, M. E.; Sandgren, M.; Stahlberg, J.; Beckham, G. T., Fungal cellulases. *Chem Rev* **2015**, 115, 1308-1448.

186. Lynd, L. R.; Weimer, P. J.; van Zyl, W. H.; Pretorius, I. S., Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol Mol Biol Rev* **2002**, 66, 506-77, table of contents.

187. Knott, B. C.; Haddad Momeni, M.; Crowley, M. F.; Mackenzie, L. F.; Gotz, A. W.; Sandgren, M.; Withers, S. G.; Stahlberg, J.; Beckham, G. T., The mechanism of cellulose hydrolysis by a two-step, retaining cellobiohydrolase elucidated by structural and transition path sampling studies. *J Am Chem Soc* **2014**, 136, 321-9.

188. Kleywegt, G. J.; Zou, J. Y.; Divne, C.; Davies, G. J.; Sinning, I.; Stahlberg, J.; Reinikainen, T.; Srisodsuk, M.; Teeri, T. T.; Jones, T. A., The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 A resolution, and a comparison with related enzymes. *J Mol Biol* **1997**, 272, 383-97.

189. Zhang, Y. H. P.; Lynd, L. R., Toward an aggregated understanding of enzymatic hydrolysis of cellulose: noncomplexed cellulase systems. *Biotech Bioeng* **2004**, 88, 797-824.

190. Bu, L.; Nimlos, M. R.; Shirts, M. R.; Stahlberg, J.; Himmel, M. E.; Crowley, M. F.; Beckham, G. T., Product binding varies dramatically between processive and nonprocessive cellulase enzymes. *J Biol Chem* **2012**, 287, 24807-13.

191. Murphy, L.; Cruys-Bagger, N.; Damgaard, H. D.; Baumann, M. J.; Olsen, S. N.; Borch, K.; Lassen, S. F.; Sweeney, M.; Tatsumi, H.; Westh, P., Origin of initial burst in activity for *Trichoderma reesei* endo-glucanases hydrolyzing insoluble cellulose. *J Biol Chem* **2012**, 287, 1252-60.

192. Wang, Y.; Zhang, S.; Song, X.; Yao, L., Cellulose chain binding free energy drives the processive move of cellulases on the cellulose surface. *Biotechnol Bioeng* **2016**, 113, 1873-80.

193. Cantarel, B. L.; Coutinho, P. M.; Rancurel, C.; Bernard, T.; Lombard, V.; Henrissat, B., The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* **2008**, 37, D233-D238.

194. Vinzant, T.; Adney, W.; Decker, S.; Baker, J.; Kinter, M.; Sherman, N.; Fox, J.; Himmel, M., Fingerprinting *Trichoderma reesei* hydrolases in a commercial cellulase preparation. *Appl Biochem Biotechnol* **2001**, 91, 99-107.

195. Martinez, D.; Berka, R. M.; Henrissat, B.; Saloheimo, M.; Arvas, M.; Baker, S. E.; Chapman, J.; Chertkov, O.; Coutinho, P. M.; Cullen, D.; Danchin, E. G.; Grigoriev, I. V.; Harris, P.; Jackson, M.; Kubicek, C. P.; Han, C. S.; Ho, I.; Luis, L. F.; de Leon, A. L.; KMagnuson, J. K.; Merino, S.; Misra, M.; Nelson, B.; Putnam, N.; Robbertse, B.; Salamov, A. A.; Schmoll, M.; Terry, A.; Thayer, N.; Westerholm-Parvinen, A.; Schoch, C. L.; Yao, J.; Barabote, R.; Nelson, M. A.; Detter, C.; Bruce, D.; Kuske, C. R.; Xie, G.; Richardson, P.; Rokhsar, D. S.; Lucas, S. M.; Rubin, E. M.; Dunn-Coleman, N.; Ward, M.; Brettin, T. S., Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. Hypocrea jecorina). *Nat Biotechnol* **2008**, 26, 553.

196. Moroz, O. V.; Maranta, M.; Shaghasi, T.; Harris, P. V.; Wilson, K. S.; Davies, G. J., The three-dimensional structure of the cellobiohydrolase Cel7A from *Aspergillus fumigatus* at 1.5 Å resolution. *Acta Crystallogr F Struct Biol Commun* **2015**, 71, 114-20.

197. Haddad Momeni, M.; Goedegebuur, F.; Hansson, H.; Karkehabadi, S.; Askarieh, G.; Mitchinson, C.; Larenas, E. A.; Ståhlberg, J.; Sandgren, M., Expression, crystal structure and cellulase activity of the thermostable cellobiohydrolase Cel7A from the fungus *Humicola grisea var. thermoidea*. *Acta Crystallogr Sect D: Biol Cyrstallogr* **2014**, 70, 2356-2366.

198. Kern, M.; McGeehan, J. E.; Streeter, S. D.; Martin, R. N.; Besser, K.; Elias, L.; Eborall, W.; Malyon, G. P.; Payne, C. M.; Himmel, M. E.; Schnorr, K.; Beckham, G. T.; Cragg, S. M.; Bruce, N. C.; McQueen-Mason, S. J., Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc Natl Acad Sci USA* **2013**, 110, 10189-94.

199. Parkkinen, T.; Koivula, A.; Vehmaanperä, J.; Rouvinen, J., Crystal structures of *Melanocarpus albomyces* cellobiohydrolase Cel7B in complex with cello-oligomers show high flexibility in the substrate binding. *Protein Sci* **2008**, 17, 1383-1394.

200. Munoz, I. G.; Ubhayasekera, W.; Henriksson, H.; Szabó, I.; Pettersson, G.; Johansson, G.; Mowbray, S. L.; Ståhlberg, J., Family 7 cellobiohydrolases from *Phanerochaete chrysosporium*: crystal structure of the catalytic module of Cel7D (CBH58) at 1.32 Å resolution and homology models of the isozymes. *J Mol Biol* **2001**, 314, 1097-1111.

201. Textor, L. C.; Colussi, F.; Silveira, R. L.; Serpa, V.; de Mello, B. L.; Muniz, J. R. C.; Squina, F. M.; Pereira Jr, N.; Skaf, M. S.; Polikarpov, I., Joint X-ray crystallographic and molecular dynamics study of cellobiohydrolase I from *Trichoderma harzianum*: deciphering the structural features of cellobiohydrolase catalytic activity. *FEBS J* **2013**, 280, 56-69.

202. Divne, C.; Stahlberg, J.; Reinikainen, T.; Ruohonen, L.; Pettersson, G.; Knowles, J.; Teeri, T. T.; Jones, T. A., The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Science* **1994**, 265, 524-528.

203. Sulzenbacher, G.; Schülein, M.; Davies, G. J., Structure of the endoglucanase I from *Fusarium oxysporum*: Native, cellobiose, and 3, 4-epoxybutyl β-D-cellobioside-inhibited forms, at 2.3 Å resolution. *Biochemistry* **1997**, 36, 5902-5911.

204. Mackenzie, L. F.; Sulzenbacher, G.; Divne, C.; Jones, T. A.; Woldike, H. F.; Schulein, M.; G, W. S.; Davies, G. J., Crystal structure of the family 7 endoglucanase I (Cel7B) from Humicola insolens at 2.2 Å resolution and identification of the catalytic nucleophile by trapping of the covalent glycosyl-enzyme intermediate. *Biochem J* **1998**, 335, 409-416.

205. Kadowaki, M. A. S.; Higasi, P.; de Godoy, M. O.; Prade, R. A.; Polikarpov, I., Biochemical and structural insights into a thermostable cellobiohydrolase from *Myceliophthora thermophila*. *FEBS J* **2018**, 285, 559-579.

206. Sonoda, M. T.; Godoy, A. S.; Pellegrini, V. O.; Kadowaki, M. A.; Nascimento, A. S.; Polikarpov, I., Structure and dynamics of *Trichoderma harzianum* Cel7B suggest molecular architecture adaptations required for a wide spectrum of activities on plant cell wall polysaccharides. *Biochim Biophys Acta Gen Subj* **2019**, 1863, 1015-1026.

207. Schiano-di-Cola, C.; Kolaczkowski, B.; Sorensen, T. H.; Christensen, S. J.; Cavaleiro, A. M.; Windahl, M. S.; Borch, K.; Morth, J. P.; Westh, P., Structural and biochemical characterization of a family 7 highly thermostable endoglucanase from the fungus *Rasamsonia emersonii*. *FEBS J* **2019**.

208. Payne, C. M.; Jiang, W.; Shirts, M. R.; Himmel, M. E.; Crowley, M. F.; Beckham, G. T., Glycoside hydrolase processivity is directly related to oligosaccharide binding free energy. *J Am Chem Soc* **2013**, 135, 18831-18839.

209. Divne, C.; Ståhlberg, J.; Teeri, T. T.; Jones, T. A., High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J Mol Biol* **1998**, 275, 309-325.

210. Ubhayasekera, W.; Muñoz, I. G.; Vasella, A.; Ståhlberg, J.; Mowbray, S. L., Structures of *Phanerochaete chrysosporium* Cel7D in complex with product and inhibitors. *FEBS J* **2005**, 272, 1952-1964.

211. Knott, B. C.; Crowley, M. F.; Himmel, M. E.; Ståhlberg, J.; Beckham, G. T., Carbohydrate–protein interactions that drive processive polysaccharide translocation in enzymes revealed from a computational study of cellobiohydrolase processivity. *J Am Chem Soc* **2014**, 136, 8810-8819.

212. Igarashi, K.; Koivula, A.; Wada, M.; Kimura, S.; Penttilä, M.; Samejima, M., High speed atomic force microscopy visualizes processive movement of *Trichoderma reesei* cellobiohydrolase I on crystalline cellulose. *J Biol Chem* **2009**, 284, 36186-36190.

213. Nakamura, A.; Tsukada, T.; Auer, S.; Furuta, T.; Wada, M.; Koivula, A.; Igarashi, K.; Samejima, M., The tryptophan residue at the active site tunnel entrance of Trichoderma reesei cellobiohydrolase Cel7A is important for initiation of degradation of crystalline cellulose. *J Biol Chem* **2013**, 288, 13503-13510.

214. Beckham, G. T.; Bomble, Y. J.; Matthews, J. F.; Taylor, C. B.; Resch, M. G.; Yarbrough, J. M.; Decker, S. R.; Bu, L.; Zhao, X.; McCabe, C.; Wohlert, J., The O-glycosylated linker from the Trichoderma reesei Family 7 cellulase is a flexible, disordered protein. *Biophys J* **2010**, 99, 3773-3781.

215. Sammond, D. W.; Payne, C. M.; Brunecky, R.; Himmel, M. E.; Crowley, M. F.; Beckham, G. T., Cellulase linkers are optimized based on domain type and function: insights from sequence analysis, biophysical measurements, and molecular simulation. *PloS one* **2012**, 7, e48615.

216. Harrison, M. J.; Nouwens, A. S.; Jardine, D. R.; Zachara, N. E.; Gooley, A. A.; Nevalainen, H.; Packer, N. H., Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of *Trichoderma reesei*. *Eur J Biochem* **1998**, 256, 119-27.

217. Amore, A.; Knott, B. C.; Supekar, N. T.; Shajahan, A.; Azadi, P.; Zhao, P.; Wells, L.; Linger, J. G.; Hobdey, S. E.; Vander Wall, T. A.; Shollenberger, T.; Yarbrough, J. M.; Tan, Z.; Crowley, M. F.; Himmel, M. E.; Decker, S. R.; Beckham, G. T.; Taylor, L. E., Distinct roles of N- and O-glycans in cellulase activity and stability. *Proc Natl Acad Sci USA* **2017**, 114, 13667-13672.

218. www.cazy.org/Carbohydrate-Binding-Modules.html Carbohydrate-Binding Module family classification.

219. Ståhlberg, J.; Johansson, G.; Pettersson, G., A new model for enzymatic hydrolysis of cellulose based on the two-domain structure of cellobiohydrolase I. *Nat Biotechnol* **1991**, 9, 286.

220. Van Tilbeurgh, H.; Tomme, P.; Claeyssens, M.; Bhikhabhai, R.; Pettersson, G., Limited proteolysis of the cellobiohydrolase I from *Trichoderma reesei*: Separation of functional domains. *FEBS Lett* **1986**, 204, 223-227.

221. Tomme, P.; van Tilbeurgh, H.; Pettersson, G.; Van Damme, J.; Vandekerckhove, J.; Knowles, J.; Teeri, T.; Claeyssens, M., Studies of the cellulolytic system of *Trichoderma reesei* QM 9414: analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur J Biochem* **1988**, 170, 575-581.

222. Reinikainen, T.; Ruohonen, L.; Nevanen, T.; Laaksonen, L.; Kraulis, P.; Jones, T. A.; Knowles, J. K.; Teeri, T. T., Investigation of the function of mutated cellulose-binding domains of *Trichoderma reesei* cellobiohydrolase I. *Proteins Struct Funct Bioinf* **1992**, 14, 475-482.

223. Le Costaouëc, T.; Pakarinen, A.; Várnai, A.; Puranen, T.; Viikari, L., The role of carbohydrate binding module (CBM) at high substrate consistency: comparison of *Trichoderma reese*i and *Thermoascus aurantiacus* Cel7A (CBHI) and Cel5A (EGII). *Bioresour Technol* **2013**, 143, 196-203.

224. Takashima, S.; Ohno, M.; Hidaka, M.; Nakamura, A.; Masaki, H.; Uozumi, T., Correlation between cellulose binding and activity of cellulose-binding domain mutants of *Humicola grisea* cellobiohydrolase 1. *FEBS Lett* **2007**, 581, 5891-5896.

225. Schiano-di-Cola, C.; Røjel, N.; Jensen, K.; Kari, J.; Sørensen, T. H.; Borch, K.; Westh, P., Systematic deletions in the cellobiohydrolase (CBH) Cel7A from the fungus *Trichoderma reesei* reveal flexible loops critical for CBH activity. *J Biol Chem* **2019**, 294, 1807-1815.

226. Alpaydin, E., *Introduction to machine learning*. MIT press: 2009.

227. Consortium, U., The universal protein resource (UniProt) in 2010. *Nucleic Acids Res* **2009**, 38, D142-D148.

228. Whisstock, J. C.; Lesk, A. M., Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **2003**, 36, 307-340.

229. Eddy, S. R., Profile hidden Markov models. *Bioinformatics* **1998**, 14, 755-763.

230. De Fonzo, V.; Aluffi-Pentini, F.; Parisi, V., Hidden Markov models in bioinformatics. *Curr Bioinform* **2007**, 2, 49-61.

231. Zhu, X.; Wu, X., Class noise vs. attribute noise: A quantitative study. *Artif Intell* **2004**, 22, 177-210.

232. Pechenizkiy, M.; Tsymbal, A.; Puuronen, S.; Pechenizkiy, O. Class noise and supervised learning in medical domains: The effect of feature extraction. In 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), 2006; IEEE: 2006; pp 708-713.

233. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H., Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, 16, 412-424.

234. Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biophys Acta Protein Struct* **1975**, 405, 442-451.

235. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A., The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, 91, 216-231.

236. He, H.; Garcia, E. A., Learning from imbalanced data. *IEE T Knowl Data En* **2008**, 1263-1284.

237. Drummond, C.; Holte, R. C. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In Workshop on learning from imbalanced datasets II, 2003; Citeseer: 2003; Vol. 11; pp 1-8.

238. Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J.; Baesens, B., An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis Support Syst* **2011**, 51, 141-154.

239. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **2019**, 1, 206-215.

240. Zhang, Y.; Yan, S.; Yao, L., A mechanistic study of *Trichoderma reesei* Cel7B catalyzed glycosidic bond cleavage. *J Phys Chem B* **2013**, 117, 8714-8722.

241. Lin, C.-T.; Lin, K.-L.; Yang, C.-H.; Chung, I.-F.; Huang, C.-D.; Yang, Y.-S., Protein metal binding residue prediction based on neural networks. *Int J Neural Syst* **2005**, 15, 71-84.

242. Payne, C. M.; Resch, M. G.; Chen, L.; Crowley, M. F.; Himmel, M. E.; Taylor, L. E., 2nd; Sandgren, M.; Stahlberg, J.; Stals, I.; Tan, Z.; Beckham, G. T., Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. *Proc Natl Acad Sci U S A* **2013**, 110, 14646-51.

243. Beckham, G. T.; Ståhlberg, J.; Knott, B. C.; Himmel, M. E.; Crowley, M. F.; Sandgren, M.; Sørlie, M.; Payne, C. M., Towards a molecular-level theory of carbohydrate processivity in glycoside hydrolases. *Curr Opin Biotechnol* **2014**, 27, 96-106.

244. Payne, C. M.; Baban, J.; Horn, S. J.; Backe, P. H.; Arvai, A. S.; Dalhus, B.; Bjørås, M.; Eijsink, V. G.; Sørlie, M.; Beckham, G. T.; Vaaje-Kolstad, G., Hallmarks of processivity in glycoside hydrolases from crystallographic and computational studies of the *Serratia marcescens* chitinases. *J Biol Chem* **2012**, 287, 36322-36330.

245. Colussi, F.; Sørensen, T. H.; Alasepp, K.; Kari, J.; Cruys-Bagger, N.; Windahl, M. S.; Olsen, J. P.; Borch, K.; Westh, P., Probing substrate interactions in the active tunnel of a catalytically deficient cellobiohydrolase (Cel7). *J Biol Chem* **2015**, 290, 2444-2454.

246. Sulzenbacher, G.; Schulein, M.; Davies, G. J., Structure of the endoglucanase I from Fusarium oxysporum: native, cellobiose, and 3,4-epoxybutyl beta-D-cellobioside-inhibited forms, at 2.3 A resolution. *Biochemistry* **1997**, 36, 5902-11.

247. Silveira, R. L.; Skaf, M. S., Concerted motions and large-scale structural fluctuations of *Trichoderma reesei* Cel7A cellobiohydrolase. *Phys Chem Chem Phys* **2018**, 20, 7498-7507.

248. Mitsuzawa, S.; Fukuura, M.; Shinkawa, S.; Kimura, K.; Furuta, T., Alanine substitution in cellobiohydrolase provides new insights into substrate threading. *Sci Rep* **2017**, 7, 16320.

249. Zong, Z.; Li, Q.; Hong, Z.; Fu, H.; Cai, W.; Chipot, C.; Jiang, H.; Zhang, D.; Chen, S.; Shao, X., Lysine Mutation of the Claw-Arm-Like Loop Accelerates Catalysis by Cellobiohydrolases. *J Am Chem Soc* **2019**, 141, 14451-14459.

250. Mulakala, C.; Reilly, P. J., *Hypocrea jecorina* (*Trichoderma reesei*) Cel7A as a molecular machine: a docking study. *Proteins Struct Funct Bioinf* **2005**, 60, 598-605.

251. GhattyVenkataKrishna, P. K.; Alekozai, E. M.; Beckham, G. T.; Schulz, R.; Crowley, M. F.; Uberbacher, E. C.; Cheng, X., Initial recognition of a cellodextrin chain in the cellulose-binding tunnel may affect cellobiohydrolase directional specificity. *Biophys J* **2013**, 104, 904-912.

252. Kari, J.; Olsen, J.; Borch, K.; Cruys-Bagger, N.; Jensen, K.; Westh, P., Kinetics of cellobiohydrolase (Cel7A) variants with lowered substrate affinity. *J Biol Chem* **2014**, 289, 32459-32468.

253. Taylor, C. B.; Payne, C. M.; Himmel, M. E.; Crowley, M. F.; McCabe, C.; Beckham, G. T., Binding site dynamics and aromatic–carbohydrate interactions in processive and non-processive family 7 glycoside hydrolases. *J Phys Chem B* **2013**, 117, 4924-4933.

254. Betts, M. J.; Russell, R. B. Amino acid properties and consequences of substitutions. In *Bioinformatics for Geneticists*; Wiley: West Sussex, 2003; Vol. 317, pp 289-314.

255. Huang, F.; Nau, W. M., A conformational flexibility scale for amino acids in peptides. *Angew Chem* **2003**, 42, 2269-2272.

256. Katoh, K.; Standley, D. M., MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **2013**, 30, 772-780.

257. Pei, J.; Kim, B. H.; Grishin, N. V., PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* **2008**, 36, 2295-300.

258. Okonechnikov, K.; Golosova, O.; Fursov, M.; team, U., Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **2012**, 28, 1166-7.

259. Doolittle, R. F., *Computer methods for macromolecular sequence analysis*. Academic Press: San Diego, 1996.

260. Robert, X.; Gouet, P., Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* **2014**, 42, W320-W324.

261. Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E., WebLogo: a sequence logo generator. *Genome Res* **2004**, 14, 1188-1190.

262. Eddy, S., HMMER user's guide. Biological sequence analysis using profile hidden Markov models. **2003**.

263. Boerjan, W.; Ralph, J.; Baucher, M., Lignin biosynthesis. *Annu Rev Plant Biol* **2003**, 54, 519-46.

264. Masai, E.; Katayama, Y.; Fukuda, M., Genetic and Biochemical Investigations on Bacterial Catabolic Pathways for Lignin-Derived Aromatic Compounds. *Bioscience, Biotechnology, and Biochemistry* **2007**, 71, 1-15.

265. Linger, J. G.; Vardon, D. R.; Guarnieri, M. T.; Karp, E. M.; Hunsinger, G. B.; Franden, M. A.; Johnson, C. W.; Chupka, G.; Strathmann, T. J.; Pienkos, P. T.; Beckham, G. T., Lignin valorization through integrated biological funneling and chemical catalysis. *Proc Natl Acad Sci U S A* **2014**, 111, 12013-8.

266. Bugg, T. D.; Rahmanpour, R., Enzymatic conversion of lignin into renewable chemicals. *Curr Opin Chem Biol* **2015**, 29, 10-7.

267. Abdelaziz, O. Y.; Brink, D. P.; Prothmann, J.; Ravi, K.; Sun, M.; García-Hidalgo, J.; Sandahl, M.; Hulteberg, C. P.; Turner, C.; Lidén, G.; Gorwa-Grauslund, M. F.,

Biological valorization of low molecular weight lignin. *Biotechnology Advances* **2016**, 34, 1318-1346.

268. Rinaldi, R.; Jastrzebski, R.; Clough, M. T.; Ralph, J.; Kennema, M.; Bruijnincx, P. C.; Weckhuysen, B. M., Paving the Way for Lignin Valorisation: Recent Advances in Bioengineering, Biorefining and Catalysis. *Angew Chem Int Ed Engl* **2016**, 55, 8164-215.

269. Schutyser, W.; Renders, T.; Van den Bosch, S.; Koelewijn, S. F.; Beckham, G. T.; Sels, B. F., Chemicals from lignin: an interplay of lignocellulose fractionation, depolymerisation, and upgrading. *Chemical Society Reviews* **2018**, 47, 852-908.

270. Ragauskas, A. J.; Beckham, G. T.; Biddy, M. J.; Chandra, R.; Chen, F.; Davis, M. F.; Davison, B. H.; Dixon, R. A.; Gilna, P.; Keller, M.; Langan, P.; Naskar, A. K.; Saddler, J. N.; Tschaplinski, T. J.; Tuskan, G. A.; Wyman, C. E., Lignin valorization: improving lignin processing in the biorefinery. *Science* **2014**, 344, 1246843.

271. Sun, Z.; Fridrich, B.; de Santi, A.; Elangovan, S.; Barta, K., Bright Side of Lignin Depolymerization: Toward New Platform Chemicals. *Chem Rev* **2018**, 118, 614-678.

272. Karlson, U.; Dwyer, D. F.; Hooper, S. W.; Moore, E. R.; Timmis, K. N.; Eltis, L. D., Two independently regulated cytochromes P-450 in a Rhodococcus rhodochrous strain that degrades 2-ethoxyphenol and 4-methoxybenzoate. *Journal of Bacteriology* **1993**, 175, 1467-1474.

273. Eltis, L. D. K., U.; Timmis, K.N., Purification and characterization of cytochrome $P450_{RR1}$ from *Rhodococcus rhodochrous*. *FEBS* **1993**, 213, 211-216.

274. Segura, A.; Bünz, P. V.; D'Argenio, D. A.; Ornston, L. N., Genetic Analysis of a Chromosomal Region Containing vanA and vanB, Genes Required for Conversion of Either Ferulate or Vanillate to Protocatechuate in Acinetobacter. *Journal of Bacteriology* **1999**, 181, 3494-3504.

275. Morawski, B.; Segura, A.; Ornston, L. N., Substrate range and genetic analysis of Acinetobacter vanillate demethylase. *Journal of Bacteriology* **2000**, 182, 1383-1389.

276. Masai, E.; Sasaki, M.; Minakawa, Y.; Abe, T.; Sonoki, T.; Miyauchi, K.; Katayama, Y.; Fukuda, M., A novel tetrahydrofolate-dependent O-demethylase gene is essential for growth of Sphingomonas paucimobilis SYK-6 with syringate. *Journal of Bacteriology* **2004**, 186, 2757-2765.

277. Yoshikata, T.; Suzuki, K.; Kamimura, N.; Namiki, M.; Hishiyama, S.; Araki, T.; Kasai, D.; Otsuka, Y.; Nakamura, M.; Fukuda, M.; Katayama, Y.; Masai, E., Three-Component O-Demethylase System Essential for Catabolism of a Lignin-Derived Biphenyl Compound in Sphingobium sp. Strain SYK-6. *Applied and Environmental Microbiology* **2014**, 80, 7142-7153.

278. Harada, A.; Kamimura, N.; Takeuchi, K.; Yu, H. Y.; Masai, E.; Senda, T., The crystal structure of a new O-demethylase from Sphingobium sp strain SYK-6. *Febs Journal* **2017**, 284, 1855-1867.

279. Kohler, A. C.; Mills, M. J. L.; Adams, P. D.; Simmons, B. A.; Sale, K. L., Structure of aryl O-demethylase offers molecular insight into a catalytic tyrosine-dependent mechanism. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, 114, E3205-E3214.

280. Bell, S. G.; Yang, W.; Tan, A. B.; Zhou, R.; Johnson, E. O.; Zhang, A.; Zhou, W.; Rao, Z.; Wong, L. L., The crystal structures of 4-methoxybenzoate bound CYP199A2 and CYP199A4: structural changes on substrate binding and the identification of an anion binding site. *Dalton Trans* **2012**, 41, 8703-14.

281. Bell, S. G.; Zhou, R.; Yang, W.; Tan, A. B.; Gentleman, A. S.; Wong, L. L.; Zhou, W., Investigation of the substrate range of CYP199A4: modification of the partition between hydroxylation and desaturation activities by substrate and protein engineering. *Chemistry* **2012**, 18, 16677-88.

282. Tumen-Velasquez, M.; Johnson, C. W.; Ahmed, A.; Dominick, G.; Fulk, E. M.; Khanna, P.; Lee, S. A.; Schmidt, A. L.; Linger, J. G.; Eiteman, M. A.; Beckham, G. T.; Neidle, E. L., Accelerating pathway evolution by increasing the gene dosage of chromosomal segments. *Proc Natl Acad Sci U S A* **2018**, 115, 7105-7110.

283. Farrow, S. C.; Facchini, P. J., Dioxygenases Catalyze O-Demethylation and O,O-Demethylenation with Widespread Roles in Benzylisoquinoline Alkaloid Metabolism in Opium Poppy. *The Journal of Biological Chemistry* **2013**, 288, 28997-29012.

284. Salvachúa, D.; Karp, E. M.; Nimlos, C. T.; Vardon, D. R.; Beckham, G. T., Towards lignin consolidated bioprocessing: simultaneous lignin depolymerization and product generation by bacteria. *Green Chemistry* **2015**, 17, 4951-4967.

285. Bolicke, S. M.; Ternes, W., Isolation and identification of oxidation products of syringol from brines and heated meat matrix. *Meat Science* **2016**, 118, 108-116.

286. Adelakun, O. E.; Kudanga, T.; Green, I. R.; le Roes-Hill, M.; Burton, S. G., Enzymatic modification of 2,6-dimethoxyphenol for the synthesis of dimers with high antioxidant capacity. *Process Biochemistry* **2012**, 47, 1926-1932.

287. Wan, Y. Y.; Du, Y. M.; Miyakoshi, T. S., Enzymatic catalysis of 2,6-dimethoxyphenol by laccases and products characterization in organic solutions. *Sci. China Ser. B-Chem.* **2008**, 51, 669-676.

288. Lee, Y.-T.; Wilson, R. F.; Rupniewski, I.; Goodin, D. B., P450cam Visits an Open Conformation in the Absence of Substrate. *Biochemistry* **2010**, 49, 3412-3419.

289. Sevrioukova, I. F.; Li, H.; Zhang, H.; Peterson, J. A.; Poulos, T. L., Structure of a cytochrome P450–redox partner electron-transfer complex. *Proceedings of the National Academy of Sciences* **1999**, 96, 1863-1868.

290. Haines, D. C.; Tomchick, D. R.; Machius, M.; Peterson, J. A., Pivotal Role of Water in the Mechanism of P450BM-3. *Biochemistry* **2001**, 40, 13456-13465.

291. Vardon, D. R.; Franden, M. A.; Johnson, C. W.; Karp, E. M.; Guarnieri, M. T.; Linger, J. G.; Salm, M. J.; Strathmann, T. J.; Beckham, G. T., Adipic acid production from lignin. *Energy & Environmental Science* **2015**, 8, 617-628.

292. Nogales, J.; Canales, A.; Jimenez-Barbero, J.; Serra, B.; Pingarron, J. M.; Garcia, J. L.; Diaz, E., Unravelling the gallic acid degradation pathway in bacteria: the gal cluster from Pseudomonas putida. *Mol Microbiol* **2011**, 79, 359-74.

293. Saeki, Y.; Nozaki, M.; Senoh, S., CLEAVAGE OF PYROGALLOL BY NON-HEME IRON-CONTAINING DIOXYGENASES. *Journal of Biological Chemistry* **1980**, 255, 8465-8471.

294. Guengerich, F. P.; Martin, M. V.; Sohl, C. D.; Cheng, Q., Measurement of cytochrome P450 and NADPH-cytochrome P450 reductase. *Nat Protoc* **2009**, 4, 1245-51.

295. Bell, S. G.; Tan, A. B.; Johnson, E. O.; Wong, L. L., Selective oxidative demethylation of veratric acid to vanillic acid by CYP199A4 from Rhodopseudomonas palustris HaA2. *Mol Biosyst* **2010**, 6, 206-14.

296. Chrastil, J. W., J.T., A sensitive colorimetric method for formaldehyde. *Anal Biochem* **1975**, 63, 202-207.

297. Knott, B. C.; Erickson, E.; Allen, M. D.; Gado, J. E.; Graham, R.; Kearns, F. L.; Pardo, I.; Topuzlu, E.; Anderson, J. J.; Austin, H. P.; Dominick, G.; Johnson, C. W.; Rorrer, N. A.; Szostkiewicz, C. J.; Copie, V.; Payne, C. M.; Woodcock, H. L.; Donohoe, B. S.; Beckham, G. T.; McGeehan, J. E., Characterization and engineering of a two-enzyme system for plastics depolymerization. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, 117, 25476-25485.

298. Law, K. L.; Morét-Ferguson, S.; Maximenko, N. A.; Proskurowski, G.; Peacock, E. E.; Hafner, J.; Reddy, C. M., Plastic accumulation in the North Atlantic subtropical gyre. *Science* **2010**, 329, 1185-1188.

299. Cózar, A.; Echevarría, F.; González-Gordillo, J. I.; Irigoien, X.; Úbeda, B.; Hernández-León, S.; Palma, Á. T.; Navarro, S.; García-de-Lomas, J.; Ruiz, A., Plastic debris in the open ocean. *Proc. Natl. Acad. Sci.* **2014**, 111, 10239-10244.

300. Jambeck, J. R.; Geyer, R.; Wilcox, C.; Siegler, T. R.; Perryman, M.; Andrady, A.; Narayan, R.; Law, K. L., Plastic waste inputs from land into the ocean. *Science* **2015**, 347, 768-771.

301. Lamb, J. B.; Willis, B. L.; Fiorenza, E. A.; Couch, C. S.; Howard, R.; Rader, D. N.; True, J. D.; Kelly, L. A.; Ahmad, A.; Jompa, J.; Harvell, C. D., Plastic waste associated with disease on coral reefs. *Science* **2018**, 359, 460-462.

302. de Souza Machado, A. A.; Kloas, W.; Zarfl, C.; Hempel, S.; Rillig, M. C., Microplastics as an emerging threat to terrestrial ecosystems. *Global Change Biol.* **2018**, 24, 1405-1416.

303. Allen, S.; Allen, D.; Phoenix, V. R.; Le Roux, G.; Jiménez, P. D.; Simonneau, A.; Binet, S.; Galop, D., Atmospheric transport and deposition of microplastics in a remote mountain catchment. *Nat. Geosci.* **2019**, 1.

304. Restrepo-Flórez, J.-M.; Bassi, A.; Thompson, M. R., Microbial degradation and deterioration of polyethylene–A review. *Intl. Biodeter. Biodegrad.* **2014**, 88, 83-90.

305. Yang, J.; Yang, Y.; Wu, W.-M.; Zhao, J.; Jiang, L., Evidence of polyethylene biodegradation by bacterial strains from the guts of plastic-eating waxworms. *Env. Sci. Tech.* **2014**, 48, 13776-13784.

306. Yang, Y.; Yang, J.; Wu, W.-M.; Zhao, J.; Song, Y.; Gao, L.; Yang, R.; Jiang, L., Biodegradation and mineralization of polystyrene by plastic-eating mealworms: part 2. Role of gut microorganisms. *Env. Sci. Tech.* **2015**, 49, 12087-12093.

307. Bombelli, P.; Howe, C. J.; Bertocchini, F., Polyethylene bio-degradation by caterpillars of the wax moth *Galleria mellonella*. *Curr. Biol.* **2017**, 27, R292-R293.

308. Dvořák, P.; Nikel, P. I.; Damborský, J.; de Lorenzo, V., Bioremediation 3.0: Engineering pollutant-removing bacteria in the times of systemic biology. *Biotechnol. Adv.* **2017**, 35, 845-866.

309. Wierckx, N.; Prieto, M. A.; Pomposiello, P.; de Lorenzo, V.; O'Connor, K.; Blank, L. M., Plastic waste as a novel substrate for industrial biotechnology. *Microb. Biotechnol.* **2015**, 8, 900-903.

310. Wei, R.; Zimmermann, W., Biocatalysis as a green route for recycling the recalcitrant plastic polyethylene terephthalate. *Microb. Biotechnol.* **2017**.

311. Wei, R.; Zimmermann, W., Microbial enzymes for the recycling of recalcitrant petroleum-based plastics: how far are we? *Microb. Biotechnol.* **2017**.

312. Araújo, R.; Silva, C.; O'Neill, A.; Micaelo, N.; Guebitz, G.; Soares, C. M.; Casal, M.; Cavaco-Paulo, A., Tailoring cutinase activity towards polyethylene terephthalate and polyamide 6,6 fibers. *J. Biotechnol.* **2007**, 128, 849-857.

313. Ronkvist, Å. M.; Xie, W.; Lu, W.; Gross, R. A., Cutinase-catalyzed hydrolysis of poly (ethylene terephthalate). *Macromolecules* **2009**, 42, 5128-5138.

314. Herrero Acero, E.; Ribitsch, D.; Steinkellner, G.; Gruber, K.; Greimel, K.; Eiteljoerg, I.; Trotscha, E.; Wei, R.; Zimmermann, W.; Zinn, M.; Cavaco-Paulo, A.; Freddi, G.; Schwab, H.; Guebitz, G., Enzymatic surface hydrolysis of PET: Effect of structural diversity on kinetic properties of cutinases from *Thermobifida*. *Macromolecules* **2011**, 44, 4632-4640.

315. Ribitsch, D.; Acero, E. H.; Greimel, K.; Eiteljoerg, I.; Trotscha, E.; Freddi, G.; Schwab, H.; Guebitz, G. M., Characterization of a new cutinase from *Thermobifida alba* for PET-surface hydrolysis. *Biocat. Biotrans.* **2012**, 30, 2-9.

316. Sulaiman, S.; Yamato, S.; Kanaya, E.; Kim, J.-J.; Koga, Y.; Takano, K.; Kanaya, S., Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. *Appl. Env. Microbiol.* **2012**, 78, 1556-1562.

317. Ribitsch, D.; Yebra, A. O.; Zitzenbacher, S.; Wu, J.; Nowitsch, S.; Steinkellner, G.; Greimel, K.; Doliska, A.; Oberdorfer, G.; Gruber, C. C., Fusion of binding domains to *Thermobifida cellulosilytica* cutinase to tune sorption characteristics and enhancing PET hydrolysis. *Biomacromolecules* **2013**, 14, 1769-1776.

318. Roth, C.; Wei, R.; Oeser, T.; Then, J.; Föllner, C.; Zimmermann, W.; Sträter, N., Structural and functional studies on a thermostable polyethylene terephthalate degrading hydrolase from *Thermobifida fusca*. *Appl. Microbiol. Biotechnol.* **2014**, 98, 7815-7823.

319. Ribitsch, D.; Acero, E. H.; Przylucka, A.; Zitzenbacher, S.; Marold, A.; Gamerith, C.; Tscheließnig, R.; Jungbauer, A.; Rennhofer, H.; Lichtenegger, H., Enhanced cutinase-catalyzed hydrolysis of polyethylene terephthalate by covalent fusion to hydrophobins. *Appl. Env. Microbiol.* **2015**, 81, 3586-3592.

320. de Castro, A. M.; Carniel, A.; Nicomedes Junior, J.; da Conceição Gomes, A.; Valoni, É., Screening of commercial enzymes for poly(ethylene terephthalate) (PET) hydrolysis and synergy studies on different substrate sources. *J. Ind. Microbiol. Biotechnol.* **2017**, 44, 835-844.

321. Han, X.; Liu, W.; Huang, J.-W.; Ma, J.; Zheng, Y.; Ko, T.-P.; Xu, L.; Cheng, Y.-S.; Chen, C.-C.; Guo, R.-T., Structural insight into catalytic mechanism of PET hydrolase. *Nature Comm.* **2017**, 8, 2106.

322. Fecker, T.; Galaz-Davison, P.; Engelberger, F.; Narui, Y.; Sotomayor, M.; Parra, L. P.; Ramírez-Sarmiento, C. A., Active site flexibility as a hallmark for efficient PET degradation by *I. sakaiensis* PETase. *Biophys. J.* **2018**, 114, 1302-1312.

323. Furukawa, M.; Kawakami, N.; Oda, K.; Miyamoto, K., Acceleration of enzymatic degradation of poly (ethylene terephthalate) by surface coating with anionic surfactants. *ChemSusChem* **2018**, 11, 4018-4025.

324. Ma, Y.; Yao, M.; Li, B.; Ding, M.; He, B.; Chen, S.; Zhou, X.; Yuan, Y., Enhanced poly(ethylene terephthalate) hydrolase activity by protein engineering. *Engineering* **2018**, 4, 888-893.

325. Son, H. F.; Cho, I. J.; Joo, S.; Seo, H.; Sagong, H.-Y.; Choi, S. Y.; Lee, S. Y.; Kim, K.-J., Rational protein engineering of thermo-stable PETase from *Ideonella sakaiensis* for highly efficient PET degradation. *ACS Catal.* **2019**, 9, 3519-3526.

326. Rauwerdink, A.; Kazlauskas, R. J., How the same core catalytic machinery catalyzes 17 different reactions: the serine-histidine-aspartate catalytic triad of α/β-hydrolase fold enzymes. *ACS Catal.* **2015**, 5, 6153-6176.

327. Hermoso, J. A.; Sanz-Aparicio, J.; Molina, R.; Juge, N.; Gonzalez, R.; Faulds, C. B., The crystal structure of feruloyl esterase A from *Aspergillus niger* suggests evolutive functional convergence in feruloyl esterase family. *J. Mol. Biol.* **2004**, 338, 495-506.

328. Suzuki, K.; Hori, A.; Kawamoto, K.; Thangudu, R. R.; Ishida, T.; Igarashi, K.; Samejima, M.; Yamada, C.; Arakawa, T.; Wakagi, T., Crystal structure of a feruloyl esterase belonging to the tannase family: a disulfide bond near a catalytic triad. *Proteins: Struct. Function Bioinform.* **2014**, 82, 2857-2867.

329. Brooks, B. R.; Brooks Iii, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S., CHARMM: the biomolecular simulation program. *J. Comp. Chem.* **2009**, 30, 1545-1614.

330. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J. Comp. Chem.* **2005**, 26, 1781-1802.

331. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *J. Chem. Theory Comput.* **2012**, 8, 3257-3273.

332. Case, D. A.; Cheatham Iii, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J. Comp. Chem.* **2005**, 26, 1668-1688.

333. Gaus, M.; Goez, A.; Elstner, M., Parametrization and benchmark of DFTB3 for organic molecules. *J. Chem. Theory Comput.* **2013**, 9, 338-354.

334. Knott, B. C.; Haddad Momeni, M.; Crowley, M. F.; Mackenzie, L. F.; Götz, A. W.; Sandgren, M.; Withers, S. G.; Ståhlberg, J.; Beckham, G. T., The mechanism of cellulose hydrolysis by a two-step, retaining cellobiohydrolase elucidated by structural and transition path sampling studies. *J. Amer. Chem. Soc.* **2013**, 136, 321-329.

335. Mayes, H. B.; Knott, B. C.; Crowley, M. F.; Broadbelt, L. J.; Ståhlberg, J.; Beckham, G. T., Who's on base? Revealing the catalytic mechanism of inverting family 6 glycoside hydrolases. *Chem. Sci.* **2016**, 7, 5955-5968.

336. Liu, J.; Hamza, A.; Zhan, C.-G., Fundamental reaction mechanism and free energy profile for (−)-cocaine hydrolysis catalyzed by cocaine esterase. *J. Amer. Chem. Soc.* **2009**, 131, 11964-11975.

337. Smith, A. J. T.; Müller, R.; Toscano, M. D.; Kast, P.; Hellinga, H. W.; Hilvert, D.; Houk, K. N., Structural reorganization and preorganization in enzyme active sites: comparisons of experimental and theoretically ideal active site geometries in the multistep serine esterase reaction cycle. *J. Amer. Chem. Soc.* **2008**, 130, 15361-15373.

338. Zhou, Y.; Wang, S.; Zhang, Y., Catalytic reaction mechanism of acetylcholinesterase determined by Born− Oppenheimer ab initio QM/MM molecular dynamics simulations. *J. Phys. Chem. B* **2010**, 114, 8817-8825.

339. InterPro Database. https://www.ebi.ac.uk/interpro/entry/IPR011118

340. Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **1997**, 25, 3389-3402.

341. Narayan, K. D.; Pandey, S. K.; Das, S. K., Characterization of *Comamonas thiooxidans* sp. nov., and comparison of thiosulfate oxidation with *Comamonas testosteroni* and *Comamonas composti*. *Curr. Microbiol.* **2010**, 61, 248-253.

342. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T., SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acid Res.* **2018**, 46, W296-W303.

343. Crepin, V. F.; Faulds, C. B.; Connerton, I. F., A non-modular type B feruloyl esterase from *Neurospora crassa* exhibits concentration-dependent substrate inhibition. *Biochem. J.* **2003**, 370, 417-427.

344. Jaeger, K.-E.; Ransac, S.; Dijkstra, B. W.; Colson, C.; van Heuvel, M.; Misset, O., Bacterial lipases. *FEMS Micro. Rev.* **1994**, 15, 29-63.

345. Chen, X.; Zaro, J. L.; Shen, W.-C., Fusion protein linkers: property, design and functionality. *Adv Drug Deliv Rev* **2013**, 65, 1357-1369.

346. Gandini, A.; Silvestre, A. J.; Neto, C. P.; Sousa, A. F.; Gomes, M., The furan counterpart of poly (ethylene terephthalate): An alternative material based on renewable resources. *J. Polymer Sci. A: Polymer Chem.* **2009**, 47, 295-298.

347. Barth, M.; Honak, A.; Oeser, T.; Wei, R.; Belisário-Ferrari, M. R.; Then, J.; Schmidt, J.; Zimmermann, W., A dual enzyme system composed of a polyester hydrolase and a carboxylesterase enhances the biocatalytic degradation of polyethylene terephthalate films. *Biotechnol. J.* **2016**, 11, 1082-1087.

348. Carniel, A.; Valoni, É.; Nicomedes, J.; Gomes, A. d. C.; Castro, A. M. d., Lipase from *Candida antarctica* (CALB) and cutinase from *Humicola insolens* act synergistically for PET hydrolysis to terephthalic acid. *Process Biochem.* **2017**, 59, 84-90.

349. Sasoh, M.; Masai, E.; Ishibashi, S.; Hara, H.; Kamimura, N.; Miyauchi, K.; Fukuda, M., Characterization of the terephthalate degradation genes of *Comamonas* sp. strain E6. *Appl. Env. Microbiol.* **2006**, 72, 1825-1832.

350. Hosaka, M.; Kamimura, N.; Toribami, S.; Mori, K.; Kasai, D.; Fukuda, M.; Masai, E., Novel tripartite aromatic acid transporter essential for terephthalate uptake in *Comamonas* sp. strain E6. *Appl. Environ. Microbiol.* **2013**, 79, 6148-6155.

351. Shigematsu, T.; Yumihara, K.; Ueda, Y.; Morimura, S.; Kida, K., Purification and gene cloning of the oxygenase component of the terephthalate 1,2-dioxygenase system from Delftia tsuruhatensis strain T7. *FEMS Microbiol. Lett.* **2003**, 220, 255-260.

352. Chain, P. S. G.; Denef, V. J.; Konstantinidis, K. T.; Vergez, L. M.; Agulló, L.; Reyes, V. L.; Hauser, L.; Córdova, M.; Gómez, L.; González, M.; Land, M.; Lao, V.; Larimer, F.; LiPuma, J. J.; Mahenthiralingam, E.; Malfatti, S. A.; Marx, C. J.; Parnell, J. J.; Ramette, A.; Richardson, P.; Seeger, M.; Smith, D.; Spilker, T.; Sul, W. J.; Tsoi, T. V.; Ulrich, L. E.; Zhulin, I. B.; Tiedje, J. M., Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *Proc. Natl. Acad. Sci.* **2006**, 103, 15280-15287.

353. Hara, H.; Eltis, L. D.; Davies, J. E.; Mohn, W. W., Transcriptomic analysis reveals a bifurcated Terephthalate degradation pathway in Rhodococcus sp. Strain RHA1. *J. Bacteriol.* **2007**, 189, 1641-1647.

354. Choi, K. Y.; Kim, D.; Sul, W. J.; Chae, J.-C.; Zylstra, G. J.; Kim, Y. M.; Kim, E., Molecular and biochemical analysis of phthalate and terephthalate degradation by Rhodococcus sp. strain DK17. *FEMS Microbiol. Lett.* **2005**, 252, 207-213.

355. Fuchs, G.; Boll, M.; Heider, J., Microbial degradation of aromatic compounds—from one strategy to four. *Nat. Rev. Microbiol.* **2011**, 9, 803.

356. Tiso, T.; Narancic, T.; Wei, R.; Pollet, E.; Beagan, N.; Schröder, K.; Honak, A.; Jiang, M.; Kenny, S. T.; Wierckx, N.; Perrin, R.; Avérous, L.; Zimmermann, W.; O'Connor, K.; Blank, L. M., Bio-upcycling of polyethylene terephthalate. *bioRxiv* **2020**, 2020.03.16.993592.

357. Eijsink, V. G.; Vaaje-Kolstad, G.; Vårum, K. M.; Horn, S. J., Towards new enzymes for biofuels: lessons from chitinase research. *Trends Biotechnol.* **2008**, 26, 228-235.

358. Quartinello, F.; Vecchiato, S.; Weinberger, S.; Kremenser, K.; Skopek, L.; Pellis, A.; Guebitz, G., Highly selective enzymatic recovery of building blocks from wool-cotton-polyester textile waste blends. *Polymers* **2018**, 10, 1107.

359. Magnin, A.; Pollet, E.; Phalip, V.; Avérous, L., Evaluation of biological degradation of polyurethanes. *Biotech. Advances* **2020**, 39, 107457.

360. Walker, R. C.; Crowley, M. F.; Case, D. A., The implementation of a fast and accurate QM/MM potential method in Amber. *J. Comp. Chem.* **2008**, 29, 1019-1031.

361. Kumar, S.; Stecher, G.; Tamura, K., MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, 33, 1870-1874.

362. Kim, J.; Terc, H.; Flowers, L.; Whiteley, M.; Peeples, T., Novel, thermostable family-13-like glycoside hydrolase fromMethanococcus jannaschii. *Folia Microbiol.* **2001**, 46, 475-481.

363. Elleuche, S.; Antranikian, G. Starch-hydrolyzing enzymes from thermophiles. In *Thermophilic Microbes in Environmental and Industrial Biotechnology*; Springer: 2013, pp 509-533.

364. Rigoldi, F.; Donini, S.; Redaelli, A.; Parisini, E.; Gautieri, A., Engineering of thermostable enzymes for industrial applications. *APL Bioeng.* **2018**, 2, 011501.

365. Vieille, C.; Zeikus, G. J., Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* **2001**, 65, 1-43.

366. Madigan, M. T.; Martinko, J.; Parker, J., In; Pearson Educacion: 2003.

367. Gromiha, M. M.; Oobatake, M.; Sarai, A., Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem* **1999**, 82, 51-67.

368. Dehouck, Y.; Folch, B.; Rooman, M., Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. *Protein Eng. Des. Sel.* **2008**, 21, 275-8.

369. Vogt, G.; Argos, P., Protein thermal stability: hydrogen bonds or internal packing? *Fold Des* **1997**, 2, S40-6.

370. Querol, E.; Perez-Pons, J. A.; Mozo-Villarias, A., Analysis of protein conformational characteristics related to thermostability. *Prot Eng Des Sel* **1996**, 9, 265-271.

371. Modarres, H. P.; Mofrad, M.; Sanati-Nezhad, A., Protein thermostability engineering. *RSC Adv* **2016**, 6, 115252-115270.

372. Kumar, S.; Tsai, C. J.; Nussinov, R., Factors enhancing protein thermostability. *Protein Eng* **2000**, 13, 179-91.

373. Trivedi, S.; Gehlot, H. S.; Rao, S. R., Protein thermostability in Archaea and Eubacteria. *Genet Mol Res* **2006**, 5, 816-27.

374. Zhou, X. X.; Wang, Y. B.; Pan, Y. J.; Li, W. F., Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids* **2008**, 34, 25-33.

375. Haney, P. J.; Badger, J. H.; Buldak, G. L.; Reich, C. I.; Woese, C. R.; Olsen, G. J., Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. *Proc Natl Acad Sci U S A* **1999**, 96, 3578-83.

376. Chakravarty, S.; Varadarajan, R., Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett* **2000**, 470, 65-9.

377. Kannan, N.; Vishveshwara, S., Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Eng* **2000**, 13, 753-61.

378. Farias, S. T.; Bonato, M. C., Preferred amino acids and thermostability. *Genet Mol Res* **2003**, 2, 383-93.

379. Zeldovich, K. B.; Berezovsky, I. N.; Shakhnovich, E. I., Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* **2007**, 3, e5.

380. Zhang, G.; Fang, B., LogitBoost classifier for discriminating thermophilic and mesophilic proteins. *J Biotechnol* **2007**, 127, 417-24.

381. Gromiha, M. M.; Suresh, M. X., Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins* **2008**, 70, 1274-9.

382. Zhang, G.; Fang, B., Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. *Protein Pept Lett* **2006**, 13, 965-70.

383. Wu, L.-C.; Lee, J.-X.; Huang, H.-D.; Liu, B.-J.; Horng, J.-T., An expert system to predict protein thermostability using decision tree. *Expert Syst Appl* **2009**, 36, 9007-9014.

384. Taylor, T. J.; Vaisman, II, Discrimination of thermophilic and mesophilic proteins. *BMC Struct Biol* **2010**, 10 Suppl 1, S5.

385. Lin, H.; Chen, W., Prediction of thermophilic proteins using feature selection technique. *J Microbiol Methods* **2011**, 84, 67-70.

386. Wang, D.; Yang, L.; Fu, Z.; Xia, J., Prediction of thermophilic protein with pseudo amino Acid composition: an approach from combined feature selection and reduction. *Protein Pept Lett* **2011**, 18, 684-9.

387. Zuo, Y. C.; Chen, W.; Fan, G. L.; Li, Q. Z., A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins. *Amino Acids* **2013**, 44, 573-80.

388. Ebrahimi, M.; Lakizadeh, A.; Agha-Golzadeh, P.; Ebrahimie, E.; Ebrahimi, M., Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. *PLoS One* **2011**, 6, e23146.

389. Nath, A.; Chaube, R.; Karthikeyan, S. Discrimination of psychrophilic and mesophilic proteins using random forest algorithm. In 2012 International Conference on Biomedical Engineering and Biotechnology, 2012; IEEE: 2012; pp 179-182.

390. Fan, G. L.; Liu, Y. L.; Wang, H., Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition. *J Theor Biol* **2016**, 407, 138-142.

391. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W., CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, 26, 680-2.

392. Li, Y.; Middaugh, C. R.; Fang, J., A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. *BMC Bioinformatics* **2010**, 11, 62.

393. Pucci, F.; Bourgeas, R.; Rooman, M., Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Sci Rep* **2016**, 6, 23257.

394. Li, F. M.; Li, Q. Z., Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids* **2008**, 34, 119-25.

395. Cambillau, C.; Claverie, J. M., Structural and genomic correlates of hyperthermostability. *J Biol Chem* **2000**, 275, 32383-6.

396. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *J Mach Learn Res* **2011**, 12, 2825-2830.

397. He, H.; Garcia, E. A., Learning from imbalanced data. *IEEE Trans Knowl Data Eng* **2008**, 1263-1284.

398. Mrabet, N. T.; Van den Broeck, A.; Van den brande, I.; Stanssens, P.; Laroche, Y.; Lambeir, A. M.; Matthijssens, G.; Jenkins, J.; Chiadmi, M.; van Tilbeurgh, H.; et al., Arginine residues as stabilizing elements in proteins. *Biochemistry* **1992**, 31, 2239-53.

399. Szilagyi, A.; Zavodszky, P., Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **2000**, 8, 493-504.

400. Goldstein, R. A., Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: insights from the quasi-chemical approximation. *Protein Sci* **2007**, 16, 1887-95.

401. Russell, R. J.; Ferguson, J. M.; Hough, D. W.; Danson, M. J.; Taylor, G. L., The crystal structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9 A resolution. *Biochemistry* **1997**, 36, 9983-94.

402. Burley, S.; Petsko, G., Amino-aromatic interactions in proteins. *FEBS Lett* **1986**, 203, 139-143.

403. Reed, C. J.; Lewis, H.; Trejo, E.; Winston, V.; Evilia, C., Protein adaptations in archaeal extremophiles. *Archaea* **2013**, 2013.

404. Feller, G., Protein stability and enzyme activity at extreme biological temperatures. *J Phys Condens Matter* **2010**, 22, 323101.

405. Georgiou, C. D., Functional Properties of Amino Acid Side Chains as Biomarkers of Extraterrestrial Life. *Astrobiology* **2018**, 18, 1479-1496.

406. Du, Q.; Wei, D.; Chou, K. C., Correlations of amino acids in proteins. *Peptides* **2003**, 24, 1863-9.

407. Gado, J. E.; Beckham, G. T.; Payne, C. M., Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *J. Chem. Inf. Model.* **2020**, 60, 4098-4107.

408. Ku, T.; Lu, P.; Chan, C.; Wang, T.; Lai, S.; Lyu, P.; Hsiao, N., Predicting melting temperature directly from protein sequences. *Comput. Biol. Chem.* **2009**, 33, 445-450.

409. Gorania, M.; Seker, H.; Haris, P. I., Predicting a protein's melting temperature from its amino acid sequence. *Conf. Proc. IEEE. Eng. Med. Biol. Soc.* **2010**, 1820-3.

410. Pucci, F.; Dhanani, M.; Dehouck, Y.; Rooman, M., Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. *PloS One* **2014**, 9.

411. Capriotti, E.; Fariselli, P.; Casadio, R., A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **2004**, 20, i63-i68.

412. Cheng, J.; Randall, A.; Baldi, P., Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **2006**, 62, 1125-1132.

413. Montanucci, L.; Fariselli, P.; Martelli, P. L.; Casadio, R., Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinformatics* **2008**, 24, i190-5.

414. Dehouck, Y.; Grosfils, A.; Folch, B.; Gilis, D.; Bogaerts, P.; Rooman, M., Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **2009**, 25, 2537-43.

415. Li, Y.; Fang, J., PROTS-RF: a robust model for predicting mutation-induced protein stability changes. *PLoS One* **2012**, 7, e47247.

416. Berliner, N.; Teyra, J.; Colak, R.; Garcia Lopez, S.; Kim, P. M., Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One* **2014**, 9, e107353.

417. Pucci, F.; Bourgeas, R.; Rooman, M., Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Sci. Rep.* **2016**, 6, 23257.

418. Cao, H.; Wang, J.; He, L.; Qi, Y.; Zhang, J. Z., DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *J. Chem. Inf. Model.* **2019**, 59, 1508-1514.

419. Alvarez-Machancoses, O.; De Andres-Galiana, E. J.; Fernandez-Martinez, J. L.; Kloczkowski, A., Robust Prediction of Single and Multiple Point Protein Mutations Stability Changes. *Biomolecules* **2020**, 10.

420. Zhang, G.; Fang, B., Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. *Protein Pept. Lett.* **2006**, 13, 965-70.

421. Wu, L.-C.; Lee, J.-X.; Huang, H.-D.; Liu, B.-J.; Horng, J.-T., An expert system to predict protein thermostability using decision tree. *Expert Syst. Appl.* **2009**, 36, 9007-9014.

422. Li, Y.; Middaugh, C. R.; Fang, J., A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. *BMC Bioinform.* **2010**, 11, 62.

423. Lin, H.; Chen, W., Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods* **2011**, 84, 67-70.

424. Jensen, D. B.; Vesth, T. C.; Hallin, P. F.; Pedersen, A. G.; Ussery, D. W., Bayesian prediction of bacterial growth temperature range based on genome sequences. *BMC Genomics* **2012**, 13 Suppl 7, S3.

425. Fan, G. L.; Liu, Y. L.; Wang, H., Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition. *J. Theor. Biol.* **2016**, 407, 138-142.

426. Li, G.; Rabe, K. S.; Nielsen, J.; Engqvist, M. K. M., Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth. Biol.* **2019**, 8, 1411-1420.

427. Nakashima, H.; Fukuchi, S.; Nishikawa, K., Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J. Biochem.* **2003**, 133, 507-13.

428. Zeldovich, K. B.; Berezovsky, I. N.; Shakhnovich, E. I., Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **2007**, 3, e5.

429. Li, G.; Zrimec, J.; Ji, B.; Geng, J.; Larsbrink, J.; Zelezniak, A.; Nielsen, J.; Engqvist, M. K. M., Performance of regression models as a function of experiment noise. *arXiv:.1912.08141* **2019**.

430. 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., iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, 34, 2499-2502.

431. Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M., Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, 16, 1315-1322.

432. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F., An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci* **2013**, 250, 113-141.

433. Elrahman, S. M. A.; Abraham, A., A review of class imbalance problem. *J. Netw. Innov. Comput.* **2013**, 1, 332-340.

434. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F., A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE T. Syst. Man Cy. C* **2012**, 42, 463-484.

435. Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. In Conference on Artificial Intelligence in Medicine in Europe, 2001; Springer: 2001; pp 63-66.

436. Stefanowski, J.; Wilk, S. Selective pre-processing of imbalanced data for improving classification performance. In International Conference on Data Warehousing and Knowledge Discovery, 2008; Springer: 2008; pp 283-292.

437. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Mute: Majority under-sampling technique. In 8th International Conference on Information, Communications & Signal Processing, 2011; IEEE: 2011; pp 1-4.

438. Seiffert, C.; Khoshgoftaar, T. M.; Van Hulse, J.; Napolitano, A., RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst., Man, Cybern. Syst.* **2010**, 40, 185-197.

439. Zhang, Y.; Zhang, D.; Mi, G.; Ma, D.; Li, G.; Guo, Y.; Li, M.; Zhu, M., Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions. *Comput. Biol. Chem.* **2012**, 36, 36-41.

440. Branco, P.; Torgo, L.; Ribeiro, R. P. MetaUtil: Meta learning for utility maximization in regression. In International Conference on Discovery Science, 2018; Springer: 2018; pp 129-143.

441. Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P., Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics* **2018**, 34, 3666-3674.

442. Sharma, A. K.; Srivastava, G. N.; Roy, A.; Sharma, V. K., ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformatics Approaches. *Front. Pharmacol.* **2017**, 8, 880.

443. Torgo, L.; Ribeiro, R. Precision and recall for regression. In International Conference on Discovery Science, 2009; Springer: 2009; pp 332-346.

444. Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D., BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **2019**, 47, D542-D549.

445. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, 12, 2825-2830.

446. Molinaro, A. M.; Simon, R.; Pfeiffer, R. M., Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **2005**, 21, 3301-7.

447. Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, 405, 442-51.

448. Gorodkin, J., Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, 28, 367-74.

449. Torgo, L.; Ribeiro, R. Utility-based regression. In European Conference on Principles of Data Mining and Knowledge Discovery, 2007; Springer: 2007; pp 597-604.

450. Menardi, G.; Torelli, N., Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Disc.* **2014**, 28, 92-122.

451. Chawla, N. V.; Lazarevic, A.; Hall, L. O.; Bowyer, K. W. SMOTEBoost: Improving prediction of the minority class in boosting. In European conference on principles of data mining and knowledge discovery, 2003; Springer: 2003; pp 107-119.

452. Wang, S.; Yao, X. Diversity analysis on imbalanced data sets by using ensemble models. In IEEE Symposium on Computational Intelligence and Data Mining, 2009; IEEE: 2009; pp 324-331.

453. Błaszczyński, J.; Deckert, M.; Stefanowski, J.; Wilk, S. Integrating selective pre-processing of imbalanced data with ivotes ensemble. In International conference on rough sets and current trends in computing, 2010; Springer: 2010; pp 148-157.

454. Buja, A.; Stuetzle, W., The effect of bagging on variance, bias, and mean squared error. *Preprint. AT&T Labs-Research* **2000**.

455. Evans, P., Scaling and assessment of data quality. *Acta Crystallographica Section D* **2006**, 62, 72-82.

456. Winter, G., xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **2010**, 43, 186-190.

457. Evans, P., An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D* **2011**, 67, 282-292.

458. Padilla, J. E.; Yeates, T. O., A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. *Acta Crystallographica Section D* **2003**, 59, 1124-1130.

459. Kabsch, W., XDS. *Acta Crystallographica Section D: Biological Crystallography* **2010**, 66, 125-132.

460. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D* **2004**, 60, 2126-2132.

461. Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L.-W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H., PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D* **2010**, 66, 213-221.

462. Terwilliger, T. C.; Grosse-Kunstleve, R. W.; Afonine, P. V.; Moriarty, N. W.; Zwart, P. H.; Hung, L.-W.; Read, R. J.; Adams, P. D., Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D* **2008**, 64, 61-69.

463. Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D., Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D* **2012**, 68, 352-367.

464. Blomfield, I. C.; McClain, M. S.; Princ, J. A.; Calie, P. J.; Eisenstein, B. I., Type 1 fimbriation and fimE mutants of Escherichia coli K-12. *Journal of bacteriology* **1991**, 173, 5298-5307.

465. Johnson, C. W.; Beckham, G. T., Aromatic catabolic pathway selection for optimal production of pyruvate and lactate from lignin. *Metab Eng* **2015**, 28, 240-247.

466. Cadoret, F.; Soscia, C.; Voulhoux, R. Gene Transfer: Transformation/Electroporation. In *Pseudomonas Methods and Protocols*, Filloux, A.; Ramos, J.-L., Eds.; Springer New York: New York, NY, 2014, pp 11-15.

467. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L., NCBI BLAST: a better web interface. *Nucleic Acids Research* **2008**, 36, W5-W9.

468. Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, 25, 1422-1423.

469. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015**, 11, 3696-713.

470. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **2006**, 25, 247-60.

471. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004**, 25, 1157-74.

472. Shahrokh, K.; Orendt, A.; Yost, G. S.; Cheatham, T. E., 3rd, Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J Comput Chem* **2012**, 33, 119-33.

473. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993**, 98, 10089-10092.

474. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J Comput Chem* **2005**, 26, 1781-802.

475. Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H., A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry* **2001**, 22, 501-508.

476. Martyna, G. J.; Tobias, D. J.; Klein, M. L., Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **1994**, 101, 4177-4189.

477. Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R., Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics* **1995**, 103, 4613-4621.

478. Kirkwood, J. G., Statistical mechanics of fluid mixtures. *Journal of Chemical Physics* **1935**, 3, 300-313.

479. Kollman, P., FREE-ENERGY CALCULATIONS - APPLICATIONS TO CHEMICAL AND BIOCHEMICAL PHENOMENA. *Chemical Reviews* **1993**, 93, 2395-2417.

480. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, 26, 1781-1802.

481. Axelsen, P. H.; Li, D. H., Improved convergence in dual-topology free energy calculations through use of harmonic restraints. *Journal of Computational Chemistry* **1998**, 19, 1278-1283.

482. Deng, Y. Q.; Roux, B., Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *Journal of Physical Chemistry B* **2009**, 113, 2234-2246.

483. Radak, B. K.; Chipot, C.; Suh, D.; Jo, S.; Jiang, W.; Phillips, J. C.; Schulten, K.; Roux, B., Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. *Journal of Chemical Theory and Computation* **2017**, 13, 5933-5944.

484. Pohorille, A.; Jarzynski, C.; Chipot, C., Good Practices in Free-Energy Calculations. *Journal of Physical Chemistry B* **2010**, 114, 10235-10253.

485. Sindhikara, D. J.; Emerson, D. J.; Roitberg, A. E., Exchange Often and Properly in Replica Exchange Molecular Dynamics. *Journal of Chemical Theory and Computation* **2010**, 6, 2804-2808.

486. Steinbrecher, T.; Joung, I.; Case, D. A., Soft-Core Potentials in Thermodynamic Integration: Comparing One- and Two-Step Transformations. *Journal of Computational Chemistry* **2011**, 32, 3253-3263.

487. M. J. Frisch, G. W. T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian 09, Revision A. 02. Gaussian. *Inc.: Wallingford, CT* **2009**.

488. Becke, A. D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, 38, 3098-3100.

489. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, 98, 5648-5652.

490. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, 37, 785-789.

491. Ribeiro, R. F.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation. *The Journal of Physical Chemistry B* **2011**, 115, 14556-14562.

492. Zhao, Y.; Truhlar, D. G., Computational characterization and modeling of buckyball tweezers: density functional study of concave-convex [small pi][three dots, centered][small pi] interactions. *Physical Chemistry Chemical Physics* **2008**, 10, 2813-2818.

493. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, 32, 1456-1465.

494. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, 132, 154104.

495. Barone, V.; Cossi, M., Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *The Journal of Physical Chemistry A* **1998**, 102, 1995-2001.

496. Cossi, M.; Rega, N.; Scalmani, G.; Barone, V., Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *Journal of Computational Chemistry* **2003**, 24, 669-681.

497. Schutz, C. N.; Warshel, A., What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Structure, Function, and Bioinformatics* **2001**, 44, 400-417.

498. Li, L.; Li, C.; Zhang, Z.; Alexov, E., On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi. *Journal of Chemical Theory and Computation* **2013**, 9, 2126-2136.

499. Legault, C., CYLview, 1.0 b, Université de Sherbrooke, Sherbrooke, Québec, Canada, 2009. *URL http://www. cylview. org (accessed February 1, 2016)*.

500. Li, H.; Poulos, T. L., The structure of the cytochrome p450BM-3 haem domain complexed with the fatty acid substrate, palmitoleic acid. *Nature Structural Biology* **1997**, 4, 140-146.

501. Dubey, K. D.; Wang, B.; Shaik, S., Molecular Dynamics and QM/MM Calculations Predict the Substrate-Induced Gating of Cytochrome P450 BM3 and the Regio- and Stereoselectivity of Fatty Acid Hydroxylation. *Journal of the American Chemical Society* **2016**, 138, 837-845.

502. Follmer, A. H.; Mahomed, M.; Goodin, D. B.; Poulos, T. L., Substrate-Dependent Allosteric Regulation in Cytochrome P450cam (CYP101A1). *Journal of the American Chemical Society* **2018**, 140, 16222-16228.

503. Prior, J. E.; Lynch, M. D.; Gill, R. T., Broad-host-range vectors for protein expression across gram negative hosts. *Biotechnology and Bioengineering* **2010**, 106, 326-332.

504. Jayakody, L. N.; Johnson, C. W.; Whitham, J. M.; Giannone, R. J.; Black, B. A.; Cleveland, N. S.; Klingeman, D. M.; Michener, W. E.; Olstad, J. L.; Vardon, D. R.; Brown,

R. C.; Brown, S. D.; Hettich, R. L.; Guss, A. M.; Beckham, G. T., Thermochemical wastewater valorization via enhanced microbial toxicity tolerance. *Energy & Environmental Science* **2018**, 11, 1625-1638.

505. Studier, F. W., Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **2005**, 41, 207-34.

506. Smith, P. K.; Krohn, R. I.; Hermanson, G. T.; Mallia, A. K.; Gartner, F. H.; Provenzano, M. D.; Fujimoto, E. K.; Goeke, N. M.; Olson, B. J.; Klenk, D. C., Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **1985**, 150, 76-85.

507. Neidhardt, F. C.; Bloch, P. L.; Smith, D. F., Culture medium for enterobacteria. *J. Bacteriol.* **1974**, 119, 736-47.

508. Kabsch, W., XDS. *Acta Cryst. D* **2010**, 66, 125-132.

509. The CCP4 suite: programs for protein crystallography. *Acta Cryst. D* **1994**, 50, 760-3.

510. McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J., Phaser crystallographic software. *J. Appl. Crystallogr.* **2007**, 40, 658-674.

511. Sheldrick, G. M., Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Cryst. D* **2010**, 66, 479-85.

512. Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H., PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst. D* **2010**, 66, 213-21.

513. Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D., Towards automated crystallographic structure refinement with phenix.refine. *Acta Cryst. D* **2012**, 68, 352-67.

514. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Cryst. D* **2004**, 60, 2126-32.

515. Engh, R. A.; Huber, R., Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Cryst. A* **1991**, 47, 392-400.

516. 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., Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)* **2007**, 23, 2947-2948.

517. Robert, X.; Gouet, P., Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acid Res.* **2014**, 42, W320-W324.

518. Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.; Toyohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K., A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science (New York, N.Y.)* **2016**, 351, 1196-9.

519. Katoh, K.; Standley, D. M., MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, 30, 772-80.

520. Capra, J. A.; Singh, M., Predicting functionally important residues from sequence conservation. *Bioinformatics (Oxford, England)* **2007**, 23, 1875-82.

521. Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* **2009**, 25, 1422-3.

522. Eddy, S. R., Profile hidden Markov models. *Bioinformatics (Oxford, England)* **1998**, 14, 755-63.

523. Kumar, S.; Stecher, G.; Tamura, K., MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, 33, 1870-4.

524. Jones, D. T.; Taylor, W. R.; Thornton, J. M., The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, 8, 275-82.

525. Saitou, N.; Nei, M., The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, 4, 406-25.

526. Schrödinger Release 2017-3: Schrödinger Suite 2017-3 Protein Preparation Wizard.

527. Schrödinger Release 2017-3: Impact, Schrödinger, LLC, New York, NY, 2016.

528. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013**, 27, 221-34.

529. Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M., Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, 26, 1752-80.

530. Schrödinger Release 2019-1: LigPrep, Schrödinger, LLC, New York, NY, 2019.

531. Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C., Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput. Aided Mol. Des.* **2010**, 24, 591-604.

532. Schrödinger Suite 2018-4 Induced Fit Docking protocol; Glide, Schrödinger, LLC, New York, NY, 2018; Prime, Schrödinger, LLC, New York, NY, 2018.

533. Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M., Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **2007**, 21, 681-91.

534. Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R. A., New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg. Med. Chem.* **2006**, 14, 3160-73.

535. Sherman, W.; Beard, H. S.; Farid, R., Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* **2006**, 67, 83-4.

536. Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R., Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, 49, 534-53.

537. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V., H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acid Res.* **2012**, 40, W537-W541.

538. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell Jr, A. D., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *J. Chem. Theory Comput.* **2012**, 8, 3257-3273.

539. Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses. *J. Chem. Theory Comput.* **2009**, 5, 2353-2370.

540. Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K.; Foster, T. J.; Jamison, F. W.; MacKerell, A. D., CHARMM Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in Polysaccharide and Carbohydrate–Protein Modeling. *J. Chem. Theory Comput.* **2011**, 7, 3162-3180.

541. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, 79, 926-935.

542. Vanommeslaeghe, K.; MacKerell, A. D., Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model* **2012**, 52, 3144-3154.

543. Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model* **2012**, 52, 3155-3168.

544. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comp. Chem.* **2010**, 31, 671-690.

545. Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D., Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comp. Chem.* **2012**, 33, 2451-2468.

546. Austin, H. P.; Allen, M. D.; Donohoe, B. S.; Rorrer, N. A.; Kearns, F. L.; Silveira, R. L.; Pollard, B. C.; Dominick, G.; Duman, R.; El Omari, K.; Mykhaylyk, V.; Wagner, A.; Michener, W. E.; Amore, A.; Skaf, M. S.; Crowley, M. F.; Thorne, A. W.; Johnson, C. W.; Woodcock, H. L.; McGeehan, J. E.; Beckham, G. T., Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci.* **2018**, 115, E4350-E4357.

547. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.* **1977**, 23, 327-341.

548. Crowley, M. F.; Williamson, M. J.; Walker, R. C., CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software. *Intl. J. Quantum Chem.* **2009**, 109, 3767-3772.

549. Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; S. Hayik; A. Roitberg; G. Seabra; J. Swails; A.W. Götz; I. Kolossváry; K.F. Wong; F. Paesani; J. Vanicek; R.M. Wolf; J. Liu; X. Wu; S.R. Brozell; T. Steinbrecher; H. Gohlke; Q. Cai; X. Ye; J. Wang; M.-J. Hsieh; G. Cui; D.R. Roe; D.H. Mathews; M.G. Seetin; R. Salomon-Ferrer; C. Sagui; V. Babin; T. Luchko; S. Gusarov; A. Kovalenko; Kollman, P. A., AMBER 12; University of California, San Francisco. **2012**.

550. Seabra, G. d. M.; Walker, R. C.; Elstner, M.; Case, D. A.; Roitberg, A. E., Implementation of the SCC-DFTB Method for Hybrid QM/MM Simulations within the Amber Molecular Dynamics Package. *J. Phys. Chem. A* **2007**, 111, 5655-5664.

551. Lee, T.-S.; Radak, B. K.; Pabis, A.; York, D. M., A new maximum likelihood approach for free energy profile construction from molecular simulations. *J. Chem. Theory Comput.* **2013**, 9, 153-164.

552. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L., Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, 53, 291-318.

553. Peters, B.; Beckham, G. T.; Trout, B. L., Extensions to the likelihood maximization approach for finding reaction coordinates. *J. Chem. Phys.* **2007**, 127, 034109.

554. Peters, B.; Trout, B. L., Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **2006**, 125, 054108.

555. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, 14, 33-38.

556. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K., MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, 35, 1547-1549.

557. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, 7, 539.

558. Chen, V. B.; Arendall, W. B., 3rd; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Cryst. D* **2010**, 66, 12-21.

Vita

Personal Information

Place of birth	Kaduna, Kaduna State, Nigeria

Education

2016 - present	Ph.D., Chemical Engineering, University of Kentucky, USA.
2009 - 2014	B. Eng., Chemical Engineering, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.

Honors and Awards

2020	Director's Award. Renewable Resources and Enabling Center, National Renewable Energy Laboratory, Golden, Colorado, USA.
2020	International HPC Summer School on HPC Challenges in Computational Sciences Award.
2017	Best Poster Presentation Award. Midwest Enzyme Chemistry Conference (MECC).
2011	Petroleum Technology Development Fund (PTDF) Local Scholarship Scheme (LSS) Award.
2010	Federal Government Scholarship Award.
2010	Chevron National University Scholarship Award.

Publications

Characterization and engineering of a two-enzyme system for plastics depolymerization. Knott BC,[1] Erickson E,[1] Allen MD,[1] **Gado JE**,[1] Graham R, Kearns FL, Pardo I, Topuzlu E, Anderson JJ, Austin HP, Dominick G, Johnson C, Rorrer NA, Szostkiewicz CJ, Copie V, Payne CM, Woodcock L, Donohoe BS, Beckham GT, McGeehan JE. *Proc. Natl. Acad. Sci. U. S. A.* 2020. 117(41). [1]Equal first-author contribution.

Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. **Gado JE**, Beckham GT, Payne CM. *J. Chem. Inf. Model.* 60(8) 2020.

Enabling microbial syringol conversion through structure-guided engineering.

Machovina MM, Mallinson SJB, Knott BC, Meyers AW, Garcia-Borràs M, Bu L, **Gado JE,** Oliver A, Schmidt GP, Hinchen DJ, Crowley MF, Johnson CW, Neidle EL, Payne CM, Houk KN, Beckham GT, McGeehan JE, DuBois JL. *Proc. Natl. Acad. Sci. U. S. A. 116(28),* 2019.