NORTHWESTERN UNIVERSITY

Metabolic Engineering and Scale-Up Strategies for Upgrading of Renewable Lignin-Derived Feedstocks by *Acinetobacter baylyi* ADP1

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemical and Biological Engineering

By
Erika Arvay

EVANSTON, ILLINOIS

June 2023

Abstract

Efficient and sustainable utilization of global resources represents a grand but achievable challenge. By leveraging biology, we can transform abundant, but recalcitrant resources like lignin to products ranging from fuel to medicine to polymers. Efforts to do so are expansive, but challenges remain, due in no small part to the difficulty in breaking down lignin and in efficiently utilizing the diverse and variable degradation products. However, nature has evolved organisms capable of diverse and efficient catabolism of numerous lignin-derived substrates. One such organism is *Acinetobacter baylyi* ADP1. As a non-model organism, ADP1 represents both a promising platform for biological transformation and a significant challenge in engineering due to the relative lack of information regarding engineering and scale-up compared to model organisms like *E. coli* and *S.* cerevisiae. However, recent developments for ADP1-specific tools alongside engineering towards synthesis of industrially relevant products has demonstrated the suitability of ADP1 as a powerful platform for lignin upgrading.

In this dissertation, I describe efforts to engineer and develop growth strategies for ADP1. The work comprises two research projects comprised of one paper each, and suggestions for future work. First, I discuss to my first major project, where I engineer ADP1 to synthesize mevalonate through expression of the heterologous mevalonate pathway. I improve mevalonate production titers by evaluating production from various lignin-derived substrates, eliminating a native, resource-competitive pathway, and implementing fed-batch cultivation. Next, I discuss growth and scale-up strategies specifically for *Acinetobacter baylyi* ADP1. I identify nutrient limitation as the primary mode of growth limitation in minimal medium and nitrogen as the most limiting nutrient, I implement a targeted nutrient feeding strategy to increase cell density, and I explore strategies to scale ADP1 growth the lab-scale bioreactors while providing adequate aeration. I then pivot to experimental validation of a computational modeling tool for predicting the identity of unknown metabolites in an ADP1 metabolomics data set. Here I develop LCMS

methodology for analysis of high priority metabolites and knock out genes predicted to be involved in the synthesis of identified metabolites. These data confirm both the accuracy of the metabolite prediction tool and the metabolic pathways involved in the synthesis of predicted metabolites.

Finally, I outline future project directions, specifically (i) enhancing oxygen transfer, (ii) overcoming dilution effects by cell recycle, (iii) utilization of complex and industrially relevant feedstocks, (iv) leveraging bioreactors to achieve finer process control and study consortia dynamics for ADP1, and (v) metabolic engineering towards enhanced cyanophycin production.

Acknowledgment

Throughout my PhD, I've been helped along the way by many people. I'd like to first thank my advisor, Dr. Keith Tyo, for his patience, analytical mind, and support over the years. I've learned much about the scientific process and research as a whole – the power of back of the envelope calculations, how to "do good science", and how to fail fast. His mentorship has made me a more thorough, analytical, and driven scientist. I'd also like to thank my committee, Dr. Ludmilla Aristilde, Dr. Erica Hartmann, and Dr. Danielle Tullman-Ercek for their guidance over the years. From the Northwestern core facility for Integrated Molecular Structure Education and Research Center (IMSERC), I'd like to thank Saman Shafaie for his analytical chemistry wisdom. It was through his guidance that I first found a fascination with analytical chemistry and instrumentation, and with his help that I somehow managed to not destroy the Tyo lab GCMS and HPLC (though at times narrowly). Thank you for all of your time and energy and apologies for all of the times you walked over to Tyo lab only to tighten a screw for 5 seconds before walking all the way back to IMSERC. I also need to thank my mentors during my internship at Demetrix Inc. – Anna Hjelmeland for being a wonderful boss and mentor and for her encouraging words, Erin Tom for her excellent mentorship, and the rest of the AC team at Demetrix.

I have too many people in Tyo lab to thank, but there are several people who have played a particularly large role in my graduate student life. To Bradley Biggs for his mentorship and continually sunny disposition - I appreciate your seemingly magical ability to troubleshoot cloning issues and your infinite patience for everyone else in lab constantly badgering you for help. To Namita Bhan for her lighthearted yet bone-dry humor and the cat walks with the only cat I've ever known who willingly walks on a leash. To Alexandra de Paz for her words of wisdom and love of tea. To Cat Majors, thank you for your sass, so, so many wine nights, and encyclopedic knowledge of the Bachelor franchise. Kevin Shebek, my labmate and neighbor, thanks for keeping me sane while I wrote this thesis after finally catching COVID 3 years into the pandemic.

To my gym buddies Jon Strutz, Jake Martin, and the Kevins (Fitzgerald and Shebek), thanks for all the hours spent picking up and putting down weights. To anyone who pet-sat for my cat, Cin, but especially to Jenny Lee Martin who holds a special place in Cin's heart and is still the only person she'll happily trade me for, I appreciate you greatly. Lastly, to the rest of Tyo lab, thank you for making work a joy through 5.5 years, one polar vortex, and a global pandemic.

Lastly, thank you to my family and friends for the endless support. I would not have gotten through the past 6 years without my partner, Jason, who has been a wellspring of support on the bad days and a source of joy and humor on the good days. Thank you for your impeccable editing skills, for showing me good music, for humbling me with your cooking skills, and for always telling me I was good enough even when I was absolutely certain I wasn't. As for friends, I owe much to many people, but particularly thank you to Walter (Jardo #1) Thavarajah for brightening my Sunday mornings with brunch for 3 years. Thank you also to Jake (Jardo #2) Martin for being a steadfast friend and completely illogical debate opponent.

And finally, thank you to the Dungeons and Daddies podcast for making me laugh in public, to the Tomate guy for the delicious food and for knowing my name because I eat there so often, to Peppercorns Kitchen and Kuni's, without whom my pandemic takeout life would have been dismal, and to Hopewell Brewing Company for the perhaps absurd amount of beer that I've consumed in the past 6 years. To end:

"Perfect is the enemy of good" – Voltaire

"And good is the enemy of done" – Jake Martin

To my cat, Cin, and my human, Jason.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

Efficiently and sustainably utilizing resources to meet societal needs is an essential engineering goal. Historically, many of these needs have been met through petrochemical means, but due to a multitude of obvious and complex reasons—climate change, geopolitical considerations, and difficulty in synthesizing chemicals using traditional methods—we have had to adapt to a rapidly shifting environment [1–3]. Luckily, there is an abundant, biomass-derived carbon source on Earth that represents an enormous, renewable reservoir suitable as a substitute for petroleum in various and fuel production processes [4]. This comes in the form of lignin, but utilization of lignin comes with its own set of challenges.

The obstacles inherent to lignin processing have lent themselves to the common saying among lignin researchers of "you can make anything from lignin except for money." Lignin is a complex, phenylpropanoid heteropolymer with a structure that is unique to its source, be it grasses, hardwoods, or other forms of plant biomass [5,6]. Due to the variability in both composition and bonding structure, processing lignin is challenging [7]. It is generated on the scale of 50 million tons per year as a by-product in lignocellulosic processing industries primarily in the form of biomass pulp [8], but the vast majority of it is burned as solid fuel in biorefinement processes [6,9]. Depolymerization can be accomplished by a variety of means including pyrolysis and alkaline pretreatment, which generates a mixture of organic molecules like *p*-hydroxybenzoic acid, ferulic acid, *p*-coumaric acid, acetate, and can also include sugars like glucose depending on of lignocellulosic feedstock [10,11]. The inherent compositional heterogeneity of lignin-derived product streams further contributes to the difficulty of utilizing processed lignin. Separation or purification of specific compounds in the product stream is generally infeasible due to high processing costs,

thus the nature of lignin demands downstream processes capable of transforming diverse and non-constant feedstocks into value-added products.

That being said, biological upgrading of lignin already occurs in soil biomes on a global scale and provides advantages over traditional chemical processing [12]. Over billions of years, evolution has created organisms capable not only of lignin depolymerization, but also of funneling the diverse set of resulting compounds through native metabolic pathways to produce biomass and bio-derived products [13]. Lignin degradation typically proceeds through the activity of laccases, oxidases, and peroxidases secreted by white rot fungi and some microbes [14]. Degradation products, including low molecular weight lignin polymers, organic acids, sugars, and alcohols are then assimilated into biomass by the breadth of soil microbes and other fungi present in the soil biome [15]. Due to this natural glut of lignin-degrading and assimilating organisms, a variety of tools exist to utilize lignin; the primary challenge lies in accessing these tools. Approaching this challenge through the lens of synthetic biology opens avenues for lignin upgrading that up until recently remained largely unexplored.

The field of synthetic biology leverages the distinct advantages of biological systems to achieve goals either not possible or extremely challenging through traditional chemical synthesis [16]. Biology possesses the ability to catalyze chemical reactions at ambient pressure and temperature, fitness and efficiency in utilizing diverse substrates due to billions of years of evolution, and the enormous range of transformations possible through native and promiscuous enzymatic activity. By engineering microbial metabolism, we can redirect flux of chemical inputs towards desired product outputs while taking advantage of the resourceful nature of biology. Due to the vast expanse of metabolic networks, the natively possible and engineerable transformations are virtually limitless.

The current focus on engineering of primarily model organisms limits biologically feasible transformations. Synthetic biology has largely focused on engineering well-characterized model

organisms due to the abundance of tools available for engineering. Though these organisms can and have been utilized to produce a multitude of industrially and medically important molecules, nonetheless there are biologically imposed limitations that engineering cannot easily overcome [17]. For example, catabolism of lignin-related substrates is particularly challenging for organisms not natively evolved to utilize the aromatic and organic acids that derive from lignin waste streams. Though engineering efforts have attempted to bridge this gap in native metabolism, a separate approach that shows great promise seeks to engineer the metabolism of non-model organisms that natively possess the pathways and physiology to utilize lignin-derived substrates [5,18–22].

Many aromatic-degrading microorganisms, including *Acinetobacter baylyi* ADP1 possess enzymatic pathways to transform a broad range of lignin-derived substrates to biomass [23–26]. ADP1 is particularly well suited to metabolic engineering purposes due to its genetic pliability and the recent development of engineering tools [27–31]. It also possesses a native proclivity towards accumulation of acetyl-CoA-derived products, natively wax esters, and engineering efforts have leveraged this extensively [32–37]. Metabolic engineering has primarily focused on enhancing native wax ester production, but synthesis of heterologous products represents a largely unexplored realm. Furthermore, efforts to reach high culture densities and scale cultivation beyond tens of milliliters are limited, despite ADP1's suitability as an industrial workhorse [33,36]. As an obligate aerobe and as an organism capable of upgrading moderately toxic substrates, strategies should address the following challenges:

- Determining biomass yields for essential nutrients to optimize supplementation of limiting nutrients.
- Feeding strategies to maximize assimilation of carbon substrates while minimizing substrate-based inhibition

- Improving oxygen transfer to reduce growth times for high density culture without requiring costly oxygen supplementation

Development of scale-up strategies addressing these challenges is essential to facilitating future industrial applications of ADP1.

1.2 Outline

The work in this dissertation is comprised of two original research papers. First, I discuss my work engineering ADP1 for the heterologous production of mevalonate from lignin-derived substrates. Next, I discuss strategies to scale up and improve growth of ADP1 using targeted nutrient supplementation and cultivation strategies. Finally, I outline potential areas for future exploration.

Specifically, Chapter 2 presents my work to engineer expression of the heterologous mevalonate pathway in ADP1 and to improve strain productivity with metabolic engineering and cultivation strategies. I present work in which I evaluated ADP1's ability to convert various lignin-derived substrates to mevalonate. I identified specific substrate conditions, namely the use of glucose as a co-substrate with aromatic acids, to improve strain growth and mevalonate production. I eliminated flux of carbon towards wax esters by knocking out a non-essential gene that directs acetyl-CoA away from mevalonate production. I also demonstrate a fed-batch feeding strategy to highlight the utility of ADP1 as a biotransformation tool to convert *p*-hydroxybenzoic acid, a lignin-derived aromatic acid, to mevalonate. This strategy achieved a gram-per-liter titer and improved mevalonate titer 7-fold over the initial strain. Though work remains to reach industrially viable productivity, this study represented a crucial step towards heterologous product synthesis in ADP1 and also served as further evidence of ADP1's particular capability for directing high flux of acetyl-CoA towards product formation. The promise as well as the obstacles identified

herein, namely genetic instability and lower culture density, serve as motivation for future research into ADP1 as a production platform.

Chapter 3 describes my work developing methods for scale-up of ADP1 growth to 1-L bioreactors. First, I explore modes of growth limitation in ADP1 and identify nutrient limitation as the primary factor limiting culture density in minimal medium. I further identify the most limiting nutrient as nitrogen and introduce a feeding strategy to increase culture density based off of targeted feeding of ammonium chloride alongside dilute feeding of other essential nutrients. Next, I leverage the finer control over process variables available in bioreactor cultivation to develop strategies for improving ADP1 culture density and increasing the growth rate. Specifically, I implement cascade aeration to improve oxygen transfer while sparging solely air. I also utilize online signals for dissolved oxygen and carbon dioxide exhaust concentration to continuously feed the carbon substrate, *p*-hydroxybenzoate, at a growth-dependent rate. During bioreactor cultivation, I identified a secondary growth limiting factor, which became apparent during cultivation with increasing feed concentrations. Initial exploration of strategies to mitigate this obstacle indicate that feeding of dilute M9 salts may hold promise for restoring growth in stagnated ADP1 cultures.

In chapter 4, I describe a collaborative project between myself and Dr. Jon Strutz. Here, we developed a computational workflow to determine the likelihood of candidate compounds in the context of a specific organism and environment. This workflow was applied to an ADP1 metabolomics data set containing unidentified peaks. Five high-confidence candidate molecules were identified, of which two were chosen for experimental validation. I engineered ADP1 mutant strains deficient in genes associated to the two top candidate metabolites and cultured these mutants alongside wild type ADP1. Analysis of these samples will be conducted via LCMS using a method adapted from Stuani *et al* [38].

We end with Chapter 5, where I make recommendations for future work, namely (1) improving oxygen uptake through expression of heterologous hemoglobin, (2) additional feeding and cultivation strategies for increasing culture density and volumetric productivity, and (3) strategies for utilizing continuous cultivation to explore ADP1's nutrient sequestration capabilities in mono- and co-culture.

Supplemental information for Chapters 2-4 can be found in Appendices A-C.

2. Engineering Acinetobacter baylyi ADP1 for mevalonate production from lignin-derived aromatic compounds

Author's Note: *This manuscript has been published* [39].

Utilization of lignin, an abundant renewable resource, is limited by its heterogenous composition and complex structure. Biological valorization of lignin provides advantages over traditional chemical processing as it occurs at ambient temperature and pressure and does not use harsh chemicals. Furthermore, the ability to biologically funnel heterogenous substrates to products eliminates the need for costly downstream processing and separation of feedstocks. However, lack of relevant metabolic networks and low tolerance to degradation products of lignin limits the application of traditional engineered model organisms. To circumvent this obstacle, we employed Acinetobacter baylyi ADP1, which natively catabolizes lignin-derived aromatic substrates through the β-ketoadipate pathway, to produce mevalonate from lignin-derived compounds. We enabled expression of the mevalonate pathway in ADP1 and validated activity in the presence of multiple lignin-derived aromatic substrates. Furthermore, by knocking out wax ester synthesis and utilizing fed-batch cultivation, we improved mevalonate titers 7.5-fold to 1014 mg/L (6.8 mM). This work establishes a foundation and provides groundwork for future efforts to engineer improved production of mevalonate and derivatives from lignin-derived aromatics using ADP1.

2.1 Introduction

Lignin is the second most abundant biomass-derived carbon source on Earth and represents a renewable reservoir of energy-dense substrate to perform green chemistry [10]. Annual worldwide production is approximately fifty-million tons, and this production is projected to increase significantly as it is a byproduct of biofuel production technology [40]. Due to the variability in both composition and bonding structure of lignin, processing is challenging [41,42]. Furthermore,

utilization of processed lignin is difficult due to the inherent compositional heterogeneity of lignin degradation products. Currently, most lignin is treated either as a waste stream or is burned as a solid fuel in biorefinement processes [42]. However, biological upgrading of lignin provides advantages over traditional chemical processing or combustion by enabling conversion of diverse lignin-derived substrates to high value products [41].

Aromatics-degrading microorganisms can catabolize the broad range of substrates found in processed lignin [12]. The substrate funneling characteristic of microbial lignin-derived aromatics metabolism allows for the utilization of diverse lignin-derived compounds without prior separation. Several microorganisms, including *Rhodococcus* and *Pseudomonas* species, can synthesize a range of products like triacylglycerols (TAGs) and polyhydroxyalkanoate (PHA) from lignin derivatives [41,43]. *Acinetobacter baylyi* ADP1 represents a promising candidate for biological valorization of lignin-derived compounds [44]. In addition to its versatile metabolism, it possesses natural competence and native homologous recombination machinery that enable rapid and targeted genomic manipulations (Barbe et al., 2004; Elliott & Neidle, 2011). Emerging tools for ADP1 have further broadened the feasible scope of engineering and enabled rapid iteration through design-build-test-learn cycles [46,47].

Leveraging the advantageous characteristics of ADP1, we engineered a strain capable of expressing the mevalonate pathway during growth on lignin-related aromatic substrates. Mevalonate is a small molecule with applications in cosmetics and as a monomer precursor to some classes of polyesters [48]. It is also a precursor to terpenoids that have applications in industries ranging from biofuels production to flavorings and fragrances [49–51]. The dedicated pathway for mevalonate synthesis requires three acetyl-CoA and two NADPH molecules, which at high flux can strain native metabolism. Thus, expression of this pathway in ADP1 provides an opportunity to study the impact of acetyl-CoA and NADPH siphoning in ADP1 as well as to identify potential obstacles to be overcome towards utilization of ADP1 as a production host. We show

mevalonate pathway activity in the presence of various lignin-derived aromatic compounds and improved productivity by eliminating the resource competitive wax ester pathway. In addition, we conducted fed-batch cultures to evaluate productivity over time. This work demonstrates ADP1 as a host for biological valorization of lignin-derived substrates to mevalonate and adds to the body of previous efforts [52–55] to synthesize industrially important products using ADP1's potential as an metabolic engineering host and indicates targets for future engineering.

2.2 Materials and Methods

2.2.1 Strains and media

All strains and plasmids used in this study may be found in Appendix A Tables A.2 – A.3. In this study wild type (WT) ADP1 was obtained from the Ellen Neidle lab (U. Georgia) and used for cloning [30,56]. The mevalonate plasmid was constructed using the pBWB162 (Addgene #140634) [46] vector and the pJBEI-6410 (Addgene #47049) [57] mevalonate pathway. The plasmid pBWB290 was used to perform genomic knock-out via SacB/Kan^R. Genomic DNA used to amplify genome homology for the genomic knock-out was isolated using a Wizard Genomic DNA Purification kit (Promega).

Initial precultures for mevalonate production were grown in LB Broth (Fisher Scientific). M9 minimal medium (1L) was prepared in sterile filtered water with 2 mM $MgSO_4$ (heptahydrate, Sigma Aldrich), 0.1 mM $CaCl_2$ (dihydrate, Sigma Aldrich), 0.18% (w/v) (10 mM) glucose solution (unless otherwise noted) (monohydrate, Acros Organics), and M9 Minimal Salts (BD Difco) (47.8 mM disodium phosphate, 22.0 mM monopotassium phosphate, 18.7 mM ammonium chloride, 8.6 mM sodium chloride). Aromatic carbon sources were supplied at 20 mM, 10 mM, 5 mM, or 2 mM as noted from 0.5M protocatechuate (PCA) (pH 7, adjusted with 10 N NaOH) (Sigma Aldrich), 0.5M *p*-hydroxybenzoate (POB) (pH 7, adjusted with 10 N NaOH) (Sigma Aldrich), 0.25 M ferulate (pH 7, adjusted with 10 N NaOH) (Sigma Aldrich), 0.5 M benzoate (pH 7, adjusted

with 10 N sodium hydroxide) (Sigma Aldrich), or 0.5 M anthranilate (pH 7, adjusted with 10 N NaOH) (Sigma Aldrich). Kanamycin was used at a working concentration of 25 mg/L for ADP1 cultures requiring antibiotic for plasmid maintenance.

2.2.1.1 General Cloning

All primers used in this study may be found in Appendix A Table A.1. Plasmids were purified using a GeneJet Plasmid Miniprep kit (Thermo Scientific). All PCRs were performed with DNA oligomer primers (10 μM) (IDT DNA) and 2X PrimeSTAR Max DNA polymerase (Takura Bio). PCR products were purified from agarose gel using a GeneJet Gel Extraction kit (Thermo Scientific). Plasmid assembly was performed via Gibson assembly using 2X Gibson Assembly Master Mix (NEB) [58].

2.2.1.2 Construction of Plasmids

The mevalonate pathway was cloned into pBWB162, which is derived from the broad host range vector pBAV1k and is capable of *E. coli* and ADP1 replication [59]. The upper mevalonate pathway (AtoB, ERG1, and HMG1) was amplified from the pJBEI-6410 (Addgene #47049) [57] plasmid using primers ECA01/02 (Table A.1). The vector backbone containing an ADP1-compatible origin, kanamycin resistance marker, *lacI*, *lacO*, and the IPTG-inducible p_{trc} promoter region was amplified from pBWB162 (Addgene #140634) [46] using primers ECA03/04. Gibson assembly was used to piece together the mevalonate pathway insert with the pBWB162 backbone to create pMev-LacI-trc (pECA03). This plasmid was sequenced then transformed into WT ADP1, as previously described [46] to generate strain ADP1 pMev-LacI-trc (ECA10).

To enable strong expression of levansucrase, *sacB* was placed under control of the p_{trc} promoter on the pBWB162 vector. The vector pBWB162 was amplified using primers BWB645/646, and *sacB* was amplified using primers BWB647/648. The gel-extracted DNA fragments were assembled using Gibson assembly to generate pBWB290.

2.2.1.3 Construction of *acr1* knock-out strains ADP1 Δ*acr1* (ECA14) and ADP1 Δ*acr1* pMev-LacI-trc (ECA15)

This knock-out was performed using SacB/Kan^R counterselection adapted from Metzgar *et al.* [60]. Genome homology of at least 500 bp flanking *acr1* was amplified from ADP1 genomic DNA using primers ECA05/06 (forward homology) and ECA09/10 (back homology). The SacB/Kan^R cassette, containing SacB downstream of the p_{trc} promoter and Kan^R, was amplified from pBWB290 using primers ECA07/08. The parts were assembled into the trc-SacB/KanR selection cassette using overlap extension PCR [61] with primers ECA05/10. The trc-SacB/KanR selection cassette was gel extracted and transformed into WT ADP1 to create ADP1Δ*acr1*::trc-SacB/Kan^R. Transformants were plated on kanamycin LB agar plates and incubated at 30°C overnight. The next day, 32 transformant colonies were patched onto kanamycin LB agar plates and sucrose LB agar plates to confirm sucrose sensitivity. These plates were incubated at ambient (~22°C) temperature for 24 – 48 hours [62], and colonies that displayed kanamycin resistance and sucrose sensitivity were selected from the kanamycin LB agar plate. These colonies were clonally purified by streaking again on kanamycin LB agar and on sucrose LB agar and growing for 24 – 48 hours at ambient temperature. Clonally pure colonies were validated for integration of the SacB/Kan cassette at the *acr1* locus using colony PCR and primers ECA15/16.

To delete trc-SacB/Kan^R from the ADP1 genome, a counterselection cassette was constructed. First, the forward and back homologies flanking *acr1* on the ADP1 genome were amplified from the ADP1 genome using primers ECA11/12 (forward homology) and ECA13/14 (back homology). Then the *acr1* counterselection cassette comprising the forward and back homologies was created using overlap extension PCR with primers ECA11/14. ADP1 Δ*acr1*::trc-SacB/Kan^R was transformed with the *acr1* counterselection cassette, and transformants were incubated for 24 – 48 hours at ambient temperature. Subsequent colonies were screened for loss of the SacB/Kan^R cassette by patching onto sucrose LB agar and kanamycin LB agar plates and

incubating for 24 – 48 hours at ambient temperature. Colonies that grew without inhibition on sucrose LB agar and did not grow on kanamycin LB agar were clonally purified by patching onto kanamycin LB agar and sucrose LB agar again and incubating overnight at ambient temperature. Colonies that grew only on sucrose LB agar were screened for loss of the original *acr1* gene via colony PCR with primers ECA15/16. Colonies with apparent deletion were sequence validated to generate ADP1 Δ*acr1* (ECA14). pMev-LacI-trc (pECA03) was transformed into ADP1 Δ*acr1* (ECA14) to generate ADP1 Δ*acr1* pMev-LacI-trc (ECA15).

2.2.2 Culture Conditions

2.2.2.1 General Culture Conditions

All ADP1 mevalonate production cultures were grown at 30°C and 250 rpm with 25 mg/L kanamycin for plasmid maintenance in 125 mL unbaffled Erlenmeyer flasks unless otherwise noted. Precultures were started by inoculating 5 mL LB medium from glycerol stocks and grown for 12 – 16 hours. Cells were then transferred to 25 mL M9 minimal medium supplemented with aromatic acids and glucose as carbon sources and grown for an additional 12 – 16 hours prior to inoculation into cultures. To inoculate M9 minimal medium cultures, a sample of preculture was centrifuged at 4000 x g and 4°C for 10 minutes. Spent supernatant was removed, cell pellets were resuspended in M9 minimal medium to an optical density of 5 and inoculated into 25 mL M9 minimal medium at a 1:100 dilution. Following growth to OD 0.6, the expression of the mevalonate pathway was induced with 1 mM IPTG.

2.2.2.2 Varied Aromatic Growth Rate Culture Conditions

For growth rate measurements of ADP1 in varied aromatic media, optical density was measured using a Synergy H1 microplate reader (BioTek) and Flat Bottom Clear Non-sterile 96-well plates (Fisherbrand). Precultures were grown initially in 5 mL LB overnight, then diluted 1:100 into 5 mL of M9 minimal medium containing 10 mM (0.18% w/v) glucose and either POB at 20

mM, or ferulate, benzoate, or anthranilate at 5 mM and allowed to grow for an additional 16 hours in 5 mL precultures. Precultures were then centrifuged as described above to pellet cells and resuspended at an OD of 0.05 in fresh medium and pipetted into wells. Each well contained 300 μL of culture. Blank wells containing sterile media were placed alongside culture wells to control for contamination as well as measure blank absorbance. Unmeasured wells were filled with water to minimize evaporation. The plate reader was held at 30°C and 250 rpm shaking for 14 hours with absorbance readings every 15 minutes.

2.2.2.3 Fed-Batch Culture Conditions

All fed-batch cultures were precultured in LB then M9 medium and inoculated at an OD of 0.05 as described above. All strains and conditions started with M9 minimal medium initially containing 20 mM POB and 10 mM glucose. After 24 hours of growth, 20 mM POB was added to the POB-fed culture, 20 mM glucose was added to the glucose-fed culture, and sterile water was added to the non-fed culture to control for dilution. An equal volume was added to each culture to prevent dilution affects from influencing observed trends. Every 24 hours cultures were sampled and analyzed for mevalonate concentration and substrate concentration before carbon addition. After 168 hours of growth, final samples were collected, and cell pellets were analyzed for truncation of the mevalonate pathway via colony PCR.

2.2.3 Analytical Methods

2.2.3.1 Sampling

All optical density readings were taken using a Synergy H1 microplate reader (BioTek) and Flat Bottom Clear Non-sterile 96-well plates (Fisherbrand). 300 μL of total volume was pipetted into each well with appropriate dilution by fresh M9 medium.

For fed-batch cultures, every 24 hours, prior to addition of the desired carbon source, 250 μL samples were collected. These samples were centrifuged immediately at maximum speed

(17,000 x *g*) and 4°C for 10 minutes. Culture supernatant was transferred to clean tubes and stored at 4°C until analysis by HPLC or GC-MS (below). The mevalonate pathway cassette from the final timepoint cell pellet was amplified by colony PCR (cPCR) using primers ECA17/18 to screen for large pathway deletions. The cPCR amplified mevalonate pathway DNA was purified via gel extraction and sequenced to identify mutations.

2.2.3.2 High Pressure Liquid Chromatography (HPLC) Sample Analysis

After centrifugation, 80 μL of supernatant sample was stored at 4°C until analysis. Samples were run on an Agilent 1200 Series HPLC equipped with a BioRad HPX-87H chromatography column, an Agilent G1315C Diode Array Detector (DAD), and an Agilent G1362A Refractive Index Detector (RID). The mobile phase was 10% (v/v) acetonitrile, 90% (v/v) 5 mM sulfuric acid (Thermo). The column was equilibrated for 1 hour at a flow rate of 0.600 mL/min and a column temperature of 60°C. The following method was used to run samples. The injection volume was 5.00 μL. Run time was 30 minutes with post-time of 1 minute. The autosampler was maintained at a temperature of 4°C and the RID at a temperature of 35°C. The DAD signal at 194 nm was used to quantify aromatic acids and gluconate, and the RID signal was used to quantify glucose. Standards were run for each aromatic acid, for gluconate, (Sigma-Aldrich) and for glucose to determine retention times and to generate standard curves.

2.2.3.3 Gas Chromatography Mass Spectrometry (GCMS) Sample Analysis

Samples were extracted into ethyl acetate for GCMS analysis using a method adapted Dueber *et al.* [63]. For each extraction, 20 μL of culture supernatant was placed into a 1.2 mL Eppendorf tube. To this, an extraction standard of 2 μL of 10 g/L β-carophyllene in ethyl acetate was added. To catalyze the lactonization of mevalonate to mevalonolactone, 10 μL of 0.5 N hydrochloric acid was added. The sample was vortexed briefly prior to the addition of 190 μL of ethyl acetate. The sample was then sealed and shaken at 1400 rpm for 5 minutes, then centrifuged at maximum speed (17,000 x *g*) at 4°C for 5 minutes. The organic (lower density)

fraction was analyzed immediately after extraction. The following method was used to analyze samples using an Agilent 7890 GC equipped with an Agilent HP-5ms Ultra Inert column (30 m x 250 μm x 0.25 μm) and an Agilent 7000 MS. The injection volume was 1 μL. Helium was used as the carrier gas. The front inlet was run on splitless mode at a temperature of 150°C and a pressure of 12.5 psi with a total flow rate of 54 mL/min and a septum purge flow of 3 mL/min. The MSD transfer line was held at 290°C. The oven was set to an initial temperature of 60°C with a ramp rate of 15°C/min to 200°C with a hold time of 1 minute. Post-run was held at 300°C for 5 minutes.

Standards of 100 mg/L β-caryophyllene were run alongside samples to normalize for variability in extraction. Mevalonolactone concentration was calculated from a standard curve of mevalonolactone in ethyl acetate. Yields were calculated from the mevalonate titer and substrate consumption as calculated from measured initial and final substrate concentrations. Carbon yields were calculated based upon the carbon atoms present in mevalonate and the respective substrates. For mixed substrate cultivations, c-mmol/c-mol yields were calculated based on the carbon yield for the combined substrates.

2.3 Results and Discussion

2.3.1 Mevalonate pathway is functionally expressed in ADP1, but not genetically stable in protocatechuate-only medium

Initial validation of pathway functionality was performed by culturing ADP1 pMev-LacI-trc (ECA10) in minimal M9 medium under three conditions. One condition was supplied with 10 mM (1.54 g/L) PCA, a metabolic intermediate in the catabolism of lignin-derived aromatic compounds, as the sole carbon source. A second condition was supplied with 2 mM (0.31 g/L) PCA and 0.2% (w/v) (11.1 mM) glucose. The last condition was supplied with 10 mM (1.54 g/L) PCA and 0.2% (w/v) (11.1 mM) glucose. Mevalonate and cell density was measured after 24 hours of cultivation in all conditions [Fig. 1B]. These data indicate that the mevalonate pathway can be expressed

functionally in ADP1, and based on the PCA-only condition, can produce mevalonate solely from an aromatic, lignin-related carbon source ($p < 0.01$ relative to no mevalonate control). Notably, cultures containing PCA as the sole carbon source accumulated very little mevalonate and grew to a lower OD in comparison to cultures supplemented with glucose. pMev-LacI-trc (pECA03) plasmids were isolated at 24 hours and sequenced for each culture. Sequencing revealed partial deletions of the mevalonate pathway present in PCA-only cultures [Fig. 1C]. These deletions occurred at various loci and impacted multiple genes in the mevalonate pathway. PCA is known as a poor substrate due to its tendency to chelate iron, which may have caused significant cellular stress [64]. Glucose appears to obviate this stress and stabilize the plasmid.

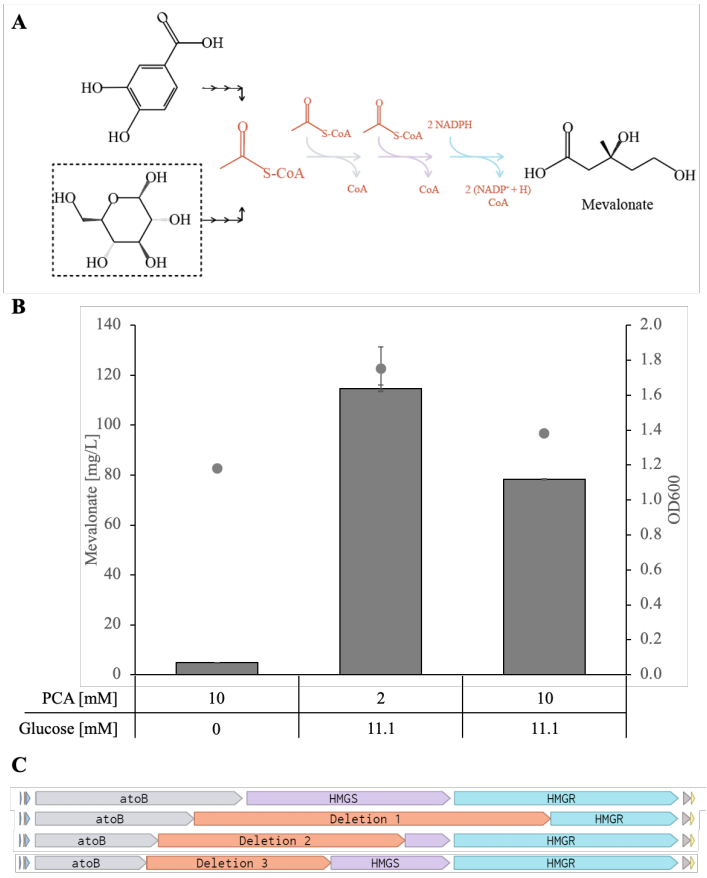

Figure 2-1 ADP1 produces mevalonate, but plasmid instability reduces production in PCA-only condition.

ADP1 was cultured in batch mode for 24 hours with expression of the mevalonate pathway. (A) The mevalonate pathway comprising *atoB, ERG13,* and *HMG1*, catalyzes the conversion of three molecules of acetyl-CoA into mevalonate. (B) Mevalonate was detected in all conditions, including in medium containing solely 10 mM PCA ($p = 0.01$, t-test, 2-tailed). The primary vertical axis shows mevalonate production. The secondary vertical axis shows OD_{600}. Data is mean and error bars represent S.E.M. ($n = 2$). (C) Plasmid instability in the 10 mM PCA condition occurred for glucose-free medium. The top band represents the full pathway, with the second, third, and fourth bands being distinct mutants found in the 10 mM PCA-only condition cultures.

2.3.2 Mevalonate production is affected by the aromatic acid species present.

After observing the sensitivity of mevalonate titers to carbon source, we examined growth and mevalonate production and yield of ECA10 in the presence of various lignin-derived aromatic acids [12]. Aromatic acid substrates were chosen to represent both branches of the β-ketoadipate pathway, funneling either to PCA (POB and ferulate) or catechol (benzoate and anthranilate) [12]. To reduce growth-inhibition effects, ferulate, benzoate, and anthranilate were supplied at 5 mM concentration, while POB was supplied at 20 mM concentration [65] [Fig. A.2]. All culture medium also contained 0.18% (w/v) (10 mM) glucose as a supplemental carbon source to ensure plasmid stability. At 48 hours, cultures were analyzed for mevalonate production and mevalonate carbon yield, as c-mmoles of mevalonate per c-mole of mixed substrate (for both aromatic acid and glucose), to account for the abundance of POB relative to other carbon sources [Fig. 2A, B]. Ferulate/glucose cultures resulted in the lowest average titer and yield (13.3 ± 1.3 mg/L and 4.9 ± 0.5 c-mmol MVA/c-mol mixed substrates, respectively), which may be due to lower tolerance of ferulate by ADP1 relative to other lignin-derived aromatic acids [54]. Analysis of carbon utilization revealed a statistically significant increase in carbon yield for POB cultures over ferulate, benzoate, and anthranilate. Cultures grown with POB/glucose had 4.6-fold higher titer (163.1 ± 11.6 mg/L) and a 2.2-fold higher yield (33.0 ± 2.4 c-mmol MVA/c-mol mixed substrates) relative to the next best performing culture, benzoate/glucose (35.5 ± 2.6 mg/L and 15.1 ± 1.1 c-mmol MVA/c-mol mixed substrates). Due to its higher mevalonate titer, yield, and low toxicity, POB/glucose was chosen as the carbon source mixture for future cultivations.

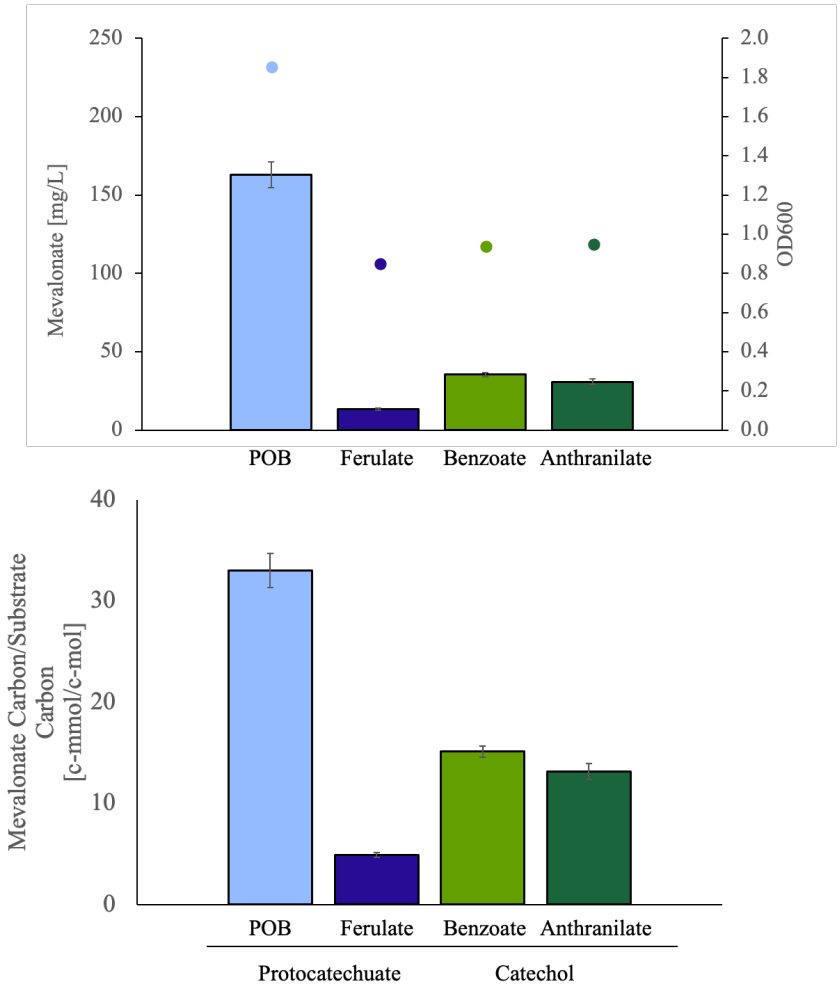

Figure 2-2 Aromatic substrates affect growth, titer, and yield.

As PCA only condition resulted in low titers and mutations, we screened four aromatics to determine which was best tolerated. Glucose was included to repress mutations found in aromatic only experiments. (A) Mevalonate titers (left axis) and optical density (right axis) after 48 h. (B) Carbon-molar yield of mevalonate on substrate. Both glucose and aromatic carbon were included as substrate. Data reflect mean, and error bars are S.E.M. (n = 4).

2.3.3 Mevalonate is efficiently produced from p-hydroxybenzoate as sole carbon source.

To evaluate the performance of cultures on solely aromatic versus aromatic and glucose medium, we conducted batch cultivation of ECA10 with M9 minimal medium containing solely 20 mM POB, solely 10 mM glucose, or a mixture of 20 mM POB and 10 mM glucose as carbon sources [Fig. A.1]. Mixed carbon substrates led to the highest titers and are more representative of the native growth conditions of ADP1 [41]. No mutations to the mevalonate plasmid were observed for any conditions. These findings were consistent under identical conditions with ADP1Δ*acr1* pMev-LacI-trc (ECA15), which lacks wax ester synthesis, thus supporting the benefit of supplying both sugar and aromatic substrates. Accordingly, glucose was included as a secondary carbon source in subsequent experiments.

2.3.4 Eliminating wax ester synthesis and culturing in fed-batch enables higher production of mevalonate from aromatic carbon

Next, we studied the impact of competition for acetyl-CoA on mevalonate production. The wax ester pathway competes for acetyl-CoA through fatty acid synthesis and is known to be highly active in ADP1 under carbon-rich and nutrient-limited, particularly nitrogen-limited, conditions [66]. Fed-batch cultivation was utilized to enable conditions that were carbon-rich and nutrient-limited. ADP1Δ*acr1* pMev-LacI-trc (ECA15) was evaluated alongside ECA10 for long-term mevalonate production in 168 hour fed-batch cultivations [Fig. 3A]. Three conditions were evaluated for mevalonate production -- one with bolus addition of POB, one with glucose, and a non-fed culture with no additional carbon subsequently added. ECA15 exhibited a significant 7.5-fold increase in mevalonate production relative to ECA10 in the POB fed-batch. For the wax ester knock-out strain expressing the mevalonate pathway, cultures reached a titer of 1014 ± 379 mg/L (6.8 mM ± 2.6 mM) with a yield of 41.4 ± 17.6 c-mmol/c-mol of all substrate consumed at 168 hrs. Glucose fed-batch also compared favorably in ECA15 relative to ECA10, although to a lesser extent than POB.

Cultures not supplied with additional non-carbon nutrients exhibited slow or insignificant growth after initial growth [Fig. A.3].

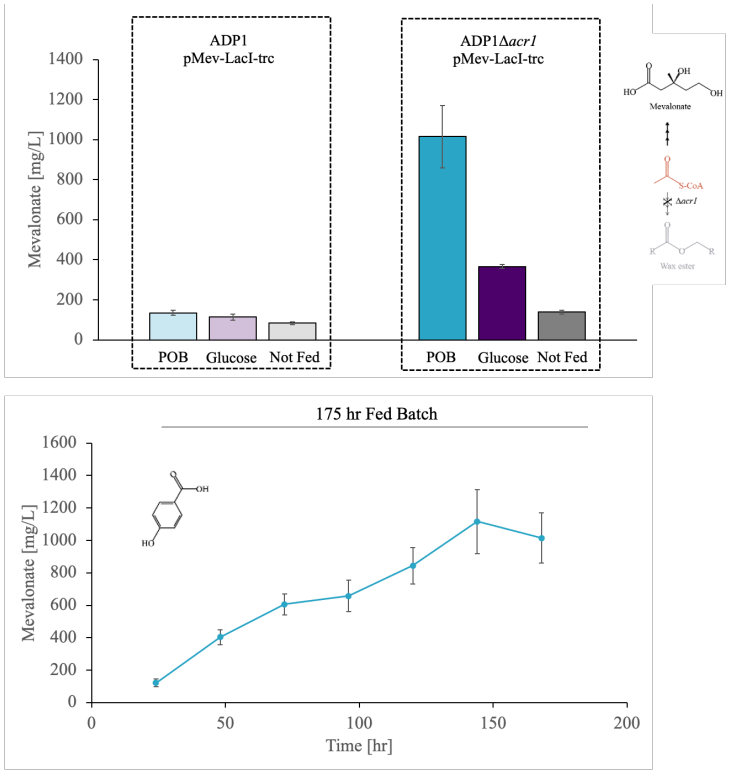

Figure 2-3 Mevalonate production is improved by wax ester knock-out and fed-batch cultivation.

A. baylyi ADP1 pMev-LacI-trc (ECA10) and the wax ester knock-out, ADP1Δ*acr1* pMev-LacI-trc (ECA15) were cultivated in fed-batch. Cultures initially contained 20 mM POB and 10 mM glucose. Cultures were fed every 24 hours one bolus of 20 mM substrate, either *p*-hydroxybenzoate (POB) or glucose, or not fed, as indicated. (A) Mevalonate titers after 168 h of POB fed-batch. (B) Mevalonate titers over culture time for in ADP1Δ*acr1* pMev-LacI-trc (ECA15) with POB feeding. Data reflect mean and error bars represent S.E.M. (n = 6 for ADP1 pMev-LacI-trc POB, glucose, not fed, and for ADP1Δ*acr1* pMev-LacI-trc POB, and n = 2 for ADP1Δ*acr1* pMev-LacI-trc glucose and not fed).

A time-course analysis of mevalonate concentration revealed that mevalonate production continued via POB catabolism even after the initial glucose was depleted [Fig. 3B, Fig. A.4 A and D]. No other C_6 compound was detected by HPLC. These data indicate that mevalonate was being produced directly from lignin-derived carbon sources over an extended timeframe. Active synthesis of mevalonate after early stationary phase and improved mevalonate synthesis in the wax ester knockout strain supports the hypothesis that the wax ester pathway competes with the mevalonate pathway for acetyl-CoA flux under carbon-rich, nutrient depleted conditions [66,67]. Amplification of the mevalonate pathway in the 168-hour cell pellets revealed truncation of the mevalonate pathway for four of twelve POB-fed cultures and for six of eight glucose-fed cultures [Table A.4, Fig. A.5]. This is consistent with the large spread in mevalonate titers late in the POB culture [Fig. 3B]. Glucose-fed cultures exhibited significant accumulation of glucose-derived gluconate that contributed to decreased pH [Table A.5] and may have reduced cell viability as well as led to mutations to the mevalonate pathway, both of which may have reduced or eliminated mevalonate production [Fig. A.4 B and E, Fig. A.5]. Therefore, we postulate that mutations are likely driven by poor cell fitness due to non-optimal pH, high substrate accumulation, or pathway expression-based burden [Fig. A.4 B and E, Fig. A.5]. Taken together, these data indicate that mevalonate production was significantly improved by eliminating the wax ester pathway, and that implementing fed-batch cultivation led to continued production of mevalonate from POB after glucose was depleted, although long-term stability remains a challenge.

2.4 Conclusions

By utilizing the diverse metabolism and genetic maleability of *Acinetobacter baylyi* ADP1, we engineered a strain capable of producing mevalonate at titers up to 1014 mg/L (6.8 mM) from mixed glucose and lignin-derived aromatic carbon sources. The strain showed functional expression of a heterologous pathway in the presence of multiple lignin-derived aromatic

compounds, and production capability was significantly improved by a genetic knock-out targeting wax ester synthesis. However, genetic instability that gradually eliminates mevalonate production poses an obstacle to industrial applications of this strain. In the future, our work will address the long-term stability of the mevalonate pathway. Chromosomal integration may enable expression while limiting plasmid-based metabolic burden, but it is unlikely to completely alleviate pathway burden. Based on our data, culture medium containing a mixture of aromatic and sugar-based carbon sources may reduce mutational frequency, indicating promise for future lignin-based medium optimization. In all, this work lays a foundation for lignin-based metabolic engineering to produce mevalonate pathway products in ADP1 and provides a case study for exploring the impact of the expression of a burdensome heterologous pathway on native ADP1 metabolism.

3. Targeted nutrient supplementation and enhanced aeration improve scale-up and growth for *Acinetobacter baylyi* ADP1

Author's Note: *This manuscript is currently in preparation.*

Industrialization of non-model organisms like *Acinetobacter baylyi* ADP1 towards establishment of sustainable biological lignin valorization processes requires investment into scale up strategies. ADP1 research efforts have primarily focused on engineering over process scale up strategies, therefore development of such strategies represents an untapped well of potential advances. Here, we demonstrate that ADP1's growth is limited primarily by nutrient limitation in minimal medium, and that nitrogen is the most limiting nutrient. Evaluation of growth on spent medium reveals that growth limitation due to small molecule inhibition is not likely. We scale ADP1 from shake flask to bench-scale bioreactors and implement targeted nutrient feeding to achieve 3.9-fold increased cell density as well as prolonging cell growth and viability. Lastly, we postulate that feeding of dilute M9 salts supplemented with additional nitrogen is sufficient to sustain growth from solely aromatic acid substrates and increase culture density of ADP1.

3.1 Introduction

Biology's ability to leverage the renewable reservoir of lignocellulosic carbon bypasses current bottlenecks in lignin processing [5,12,68]. Untapped potential exists in endogenous metabolic pathways commonly found in non-model organisms. The non-model organism *Acinetobacter baylyi* ADP1's wealth of biotransformation capability [28,69,70] is a promising tool for upgrading of lignocellulosic substrates, but effective utilization demands the development of industrialization strategies. Efficiently attaining high cellular densities has not been a primary focus of ADP1 product synthesis and engineering studies, thus exploring this area has the potential to yield significant and universal gains in production titers. Here we develop strategies for high cell density

ADP1 cultivation to demonstrate its potential as an alternative industrial chassis to traditional chemical processing.

ADP1's genetic tractability and native capacity for funneling a broad range of renewable lignin-derived substrates to central metabolism make it an ideal candidate for metabolic transformations of lignin waste streams [28,30,60,69,71]. It possesses high tolerance for lignin-derived substrates, which inhibit growth of model organisms such as *E. coli* and *S. cerevisiae* [21,72,73]. This trait particularly lends itself to biological upgrading of lignocellulosic waste streams and capitalizes on substrate-based inhibition of potential bacterial contaminants [74]. Recent developments in ADP1-specific toolsets have further enabled metabolic engineering efforts towards synthesizing products from cheap and renewable feedstocks [75,76]. However, process scale-up, which is a major hurdle in industrialization, remains largely unexplored for ADP1.

Efforts to improve product synthesis in ADP1, especially in the upgrading of lignocellulosic-derived substrates, have primarily focused on metabolic engineering and product synthesis over scale-up strategies [32,39,77]. ADP1's native proclivity for transforming a broad array of sugar and aromatic substrates to acetyl-CoA is especially promising for synthesis of products derived from TCA cycle intermediates [32,33,67,77,78]. ADP1 has been engineered for improved production of triacylglycerol [53] and wax esters [32,77,79], and for the heterologous production of 1-alkenes [80] and mevalonate [39]. While these efforts demonstrate ADP1's utility, they forego granularity in scale-up strategy for the sake of metabolic engineering.

Process scale-up strategies typically focus on improving oxygen and mass transfer, fine-tuning substrate and nutrient supplementation, and avoiding accumulation of inhibitory compounds [81–84]. These considerations are critical in high density growth. Oxygen and substrate gradients can slow growth, and depletion of nutrients or accumulation of inhibitory metabolites can reduce cell viability [83,85]. The impact of nutrient feeding strategies and industrial stressors

such as oxygen limitation and agitation-related shear is largely unexplored in ADP1. Nutrient requirements are particularly of interest due to the preference of ADP1 for low nitrogen-to-carbon ratios during synthesis of acetyl-CoA derived products such as wax esters [77].

Currently, the maximum culture density seen in ADP1 is OD 12 in shake flask cultivation [33,77] and OD 5.3 in bioreactor cultivation [67]. To grow to high density, ADP1 is typically cultured in media containing complex nutrients like casein amino acids and yeast extract alongside high amounts of non-aromatic substrates [33,77]. For bioreactor cultivation, pure oxygen supplementation has been used to maintain high growth rates [67,77], though improved oxygenation through agitation cascade has been demonstrated at lower optical densities [33].

Here we present a scale-up strategy for ADP1 to rapidly reach high culture density in minimal medium with lignin-derived substrates and targeted nutrient feeding. We explore factors limiting growth of ADP1 on M9 minimal media and find that the primary growth limitation arises from nutrient depletion. We identify nitrogen as the primary limiting nutrient and address depletion with targeted nutrient supplementation. Lastly, we implement cascade agitation to overcome oxygen demands without the need for costly supplementation of pure oxygen even at high culture densities, which is crucial to reducing process costs on an industrial scale. With these strategies, ADP1 reached a density of 15.88 in 23 hours of cultivation with a feed comprised of a lignin-derived aromatic carbon source and ammonium chloride. These findings will serve as a foundation for future work with ADP1 towards ultimately demonstrating this strains utility as a powerful industrial biotransformation tool.

3.2 Materials and Methods

3.2.1 Cultivation Methodology

Strain and Media

Wild type *A. baylyi* ADP1 was used for cell culture studies. M9 medium contained 2 mM magnesium sulfate, 0.1 mM calcium chloride, and M9 Minimal Salts (BD Difco) (47.8 mM disodium phosphate, 22.0 mM monopotassium phosphate, 18.7 mM ammonium chloride, 8.6 mM sodium chloride). Modified bioreactor M9 medium contained 2 mM magnesium sulfate, 0.1 mM calcium chloride, 30 µM iron III chloride, 15.7 mM disodium phosphate, 12.5 mM monopotassium phosphate, and 18.7 mM ammonium chloride. Where noted, trace metals solution was provided (2 mg/L $ZnSO_4 \cdot 7H_2O$; 5 mg/L $FeSO_4 \cdot 7H_2O$; 0.2 mg/L $Na_2MoO_4 \cdot 2H_2O$; 0.2 mg/L $CuSO_4 \cdot 5H_2O$; 0.4 mg/L $CoCl_2 \cdot 6H_2O$; and 1 mg/L $MnCl_2 \cdot 2H_2O$) [33,74]. Carbon sources, unless otherwise listed, were 10 mM glucose and 20 mM *p*-hydroxybenzoate (POB) due to ADP1's tolerance for POB and improved growth on mixed glucose-POB medium [39].

Precultures

Precultures were initially cultivated in LB medium by inoculating 5 mL of LB in a 14 mL Falcon tube with a single colony and grown overnight at 30°C and 250 rpm at 45 °. They were then diluted 1:100 into the relevant minimal medium and grown for an additional 12-18 hours. To inoculate cultures, precultures were pelleted at 6800 x g and 4°C and resuspended to an OD of 5 with fresh minimal medium. Cultures were inoculated to an initial OD of 0.05 and grown at 30°C and 250 rpm unless otherwise noted.

Resuspension Cultures

Resuspension cultures were grown in baffled 250 mL shake flasks with 25 mL of M9 minimal medium supplemented with 20 mM POB and 10 mM glucose. Every 24 hours, the entire culture was centrifuged at 6800 x g and 4°C for 10 minutes to pellet cells. For high density cultures (OD greater than 6), cultures were centrifuged for 20 minutes to fully pellet cells. The cell pellets were resuspended by gently pipetting in 25 mL of fresh medium and placed into new 250 mL baffled flasks to minimize the risk of contamination and carry-over of inhibitory compounds. The spent supernatant was collected for timepoints 24, 72, and 120 hours to evaluate the presence of

inhibitory compounds in spent medium. Samples were centrifuged at maximum speed and 4°C, filtered using a 0.2 mm cellulose acetate filter, and then stored at -20°C until analysis.

Spent Medium Inhibition

The spent supernatant collected from the resuspension cultures at 24, 72, and 120 hours was split into 3 treatment groups per timepoint. To the first group, 20 mM POB, 10 mM glucose, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, and 18.7 mM NH_4Cl were added to replenish spent nutrients. Buffering salts and sodium chloride were excluded as they were assumed to not be limiting to ADP1 growth. To the second group, 1X M9 salts were added alongside 20 mM POB, 10 mM glucose, 2 mM $MgSO_4$, and 0.1 mM $CaCl_2$. To the third group, no nutrients were added. Wild type ADP1 was inoculated into (i) spent medium with added nutrients and POB/glucose, (ii) spent medium with 1X M9 and POB/glucose, (iii) spent medium, and (iv) fresh M9 minimal medium. Cultures were conducted in 5 mL of medium in culture tubes and grown for 48 hours.

Nutrient Elimination

Nutrient elimination cultures were grown in 125 mL unbaffled flasks with 25 mL of medium. Culture treatments for each timepoint are in Table 1. The initial growth for all culture conditions occurred in M9 minimal medium with 20 mM POB and 10 mM glucose. At 24 hours, cultures were supplemented with a nutrient bolus containing 23.9 mM disodium phosphate, 11.0 mM monopotassium phosphate, 9.3 mM ammonium chloride, 4.3 mM sodium chloride, 1 mM magnesium sulfate, and 0.05 mM calcium chloride with a single eliminated nutrient per condition. The positive control was supplemented with a bolus containing all nutrients. The negative control was given a bolus of sterile water. All cultures were fed a bolus of carbon substrates (20 mM POB and 10 mM glucose) at 24, 48 and 96 hours to avoid carbon limitation. Every 24 hours, cultures were pH adjusted to pH 7 using 10M hydrochloric acid. At 96 hours, cultures were supplemented with a bolus containing only the eliminated nutrient. Cultures were monitored during 24 to 96 hours for reduced growth phenotype and during 96 to 144 hours for recovered growth phenotype.

Bioreactor Cultivation

1L bioreactor (Sartorius BioStat Q+) cultivations were conducted with 0.5 L of M9 medium, 1 VVM or 3 VVM sparging with air, temperature automatically controlled to 30°C and agitation rate either set to 250 rpm (initial baseline and 0.25M POB nutrient fed cultivations) or set to 200 – 1200 rpm (improved oxygen cultures) to maintain dO_2 at or above 20%. Culture pH was automatically controlled to 7 with the addition of 2M sodium hydroxide or 2M hydrochloric acid. To control foaming, 1:100 dilution of antifoam B was added as necessary. For baseline cultivations, 0.25M POB was drip-fed to maintain a concentration near zero. For nutrient-fed cultures, 0.25M, 0.5M, or 0.72M POB was drip-fed. The feed rate was adjusted based on online signals for dissolved oxygen and exhaust gas carbon dioxide as well as offline HPLC analysis to confirm POB concentration was maintained at < 10 mM throughout cultivation. For the nutrient fed conditions with bolus nutrient feeding, a nutrient feed containing (15.7 mM disodium phosphate, 12.5 mM monopotassium phosphate, 186.9 mM ammonium chloride) was fed at a dilution of 1:100 and a micronutrient feed containing (100 mM magnesium sulfate and 5 mM calcium chloride) was fed at a dilution of 1:100 every 24 - 48 hours. For nutrient fed conditions with combined nitrogen and POB feed, the feed contained 200 mM ammonium chloride and 0.72M POB and was supplemented at a rate sufficient to maintain POB concentration in the medium < 10 mM throughout cultivation.

3.2.2 Analytical Methodology

Substrate consumption was tracked via HPLC (Agilent 1200 series) equipped with a RID and a DAD. Samples were run on a BioRad Aminex 87H column with a mobile phase composed of 90% 5 mM sulfuric acid and 10% acetonitrile at a flow rate of 0.6 mL/min. The column chamber was held at 60°C. The following signals were used for each tracked compound: RID – glucose; DAD 194 nm – gluconate; DAD 206 nm – acetate, *p*-hydroxybenzoate, and protocatechuate.

Nitrogen consumption was evaluated by measuring ammonium levels in culture supernatant using the Sigma Aldrich MAK310 Ammonia Assay kit. pH, dO_2, exhaust carbon dioxide, temperature, and agitation rate were monitored online.

3.3 Results

3.3.1 The primary growth limitation mode of ADP1 in minimal medium is nutrient limitation

We evaluated ADP1 growth with serial resuspension in fresh medium to evaluate if, at higher cell density, any non-medium related (e.g. cell-density-dependent) effects were present and impacted growth. Growth in ADP1 is likely limited by (i) nutrient limitation, (ii) presence of a toxic by-product, or (iii) a non-toxic by-product or quorum sensing molecule [Fig. 3-1A]. The growth of ADP1 was observed over 24-hour growth-resuspension cycles. During each cycle, whole cultures were centrifuged to pellet cells then resuspended in fresh medium and allowed to grow for 24 hours [Fig. 3-1B]. Serially resuspending ADP1 cultures in fresh medium enabled prolonged growth with a general linear trend in cell density and consistent increases in cell density per growth cycle [Fig. 3-1C]. Analysis of substrate assimilation via HPLC indicated that carbon sources were fully consumed during each 24-hour growth cycle. Variability in culture density and in cell density change per cycle increased for later timepoints, and the overall change in cell density per cycle was diminished for the final 96-to-120-hour cycle. Growth during these cycles resembled nutrient-limited growth, but results were not conclusive particularly at higher culture densities, for which the change in cell density per 24-hour cycle was reduced.

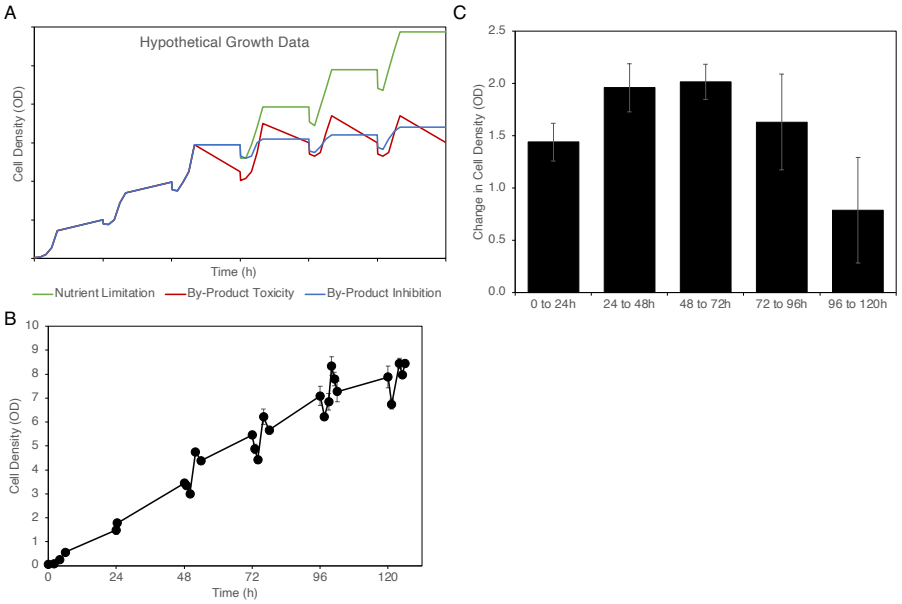

Figure 3-1 Serial resuspension of ADP1 whole culture cell pellets results in consistent and linear gains in biomass for each resuspension-growth cycle.

(A) ADP1 biomass gains for each serial whole cell pellet resuspension-growth cycle. (A) Theoretical data for resuspension-growth cycles for (i) nutrient limited growth, (ii) growth inhibition by a toxic by-product, or (iii) growth inhibition by a non-toxic by-product or density-dependent quorum sensing mechanism. (B) Time-course culture density (OD_{600}) for ADP1 WT cultures undergoing resuspension in fresh medium every 24 hours. (C) Change in cell density over each 24-hour resuspension-growth cycle. N = 4 biological replicates. Error bars are S.E.M.

3.3.2 ADP1 spent medium does not contain growth-inhibitory compounds

To evaluate if spent medium from the serially resuspended cultures contained any growth-inhibitory compounds, we collected the spent medium supernatant at 24, 72, and 120h after pelleting cells, added back essential nutrients, and inoculated fresh ADP1 cultures into the spent medium. When grown in spent medium with added nutrients and carbon, ADP1 reached a cellular density similar to that reached in the fresh medium condition at 24 hours [Fig. 3-2B]. We

hypothesized that if non-medium-dependent inhibitory effects were present and impacted growth, the effect could be more pronounced with higher culture density at later timepoints if the inhibition was strictly density-dependent. However, there was no discernable relationship between the timepoint of spent supernatant and the cell density. Growth of ADP1 on spent medium with no additional nutrients was minimal likely due to the lack of remaining carbon in the spent medium [Fig. 3-2A]. A statistically significant decrease in cell density between 24 and 48h occurred for the nutrient supplemented spent medium conditions, but not for the fresh medium condition. The pH was measured at 48 hours using pH paper and found to be between 7 and 8 for spent medium conditions and approximately 7 for the fresh medium condition. [Suppl. Table B-1]. To evaluate the impact of improved buffering capacity on growth in spent medium, 1X M9 salts were added to the spent medium alongside nutrients and carbon (2 mM magnesium sulfate, 0.1 mM calcium chloride, 20 mM POB, and 10 mM glucose). Cultures grown in spent medium with added nutrients, carbon, and M9 salts were found to reach a density similar to that of cultures in fresh medium and in spent medium with added nutrients at 24 hours [Fig. 3-2C]. The density of cultures in spent medium with added M9 salts was similar to that of cultures in fresh medium at 48 hours, indicating the inclusion of carbon, nutrients, and buffering salts can support and maintain growth for ADP1 in spent medium.

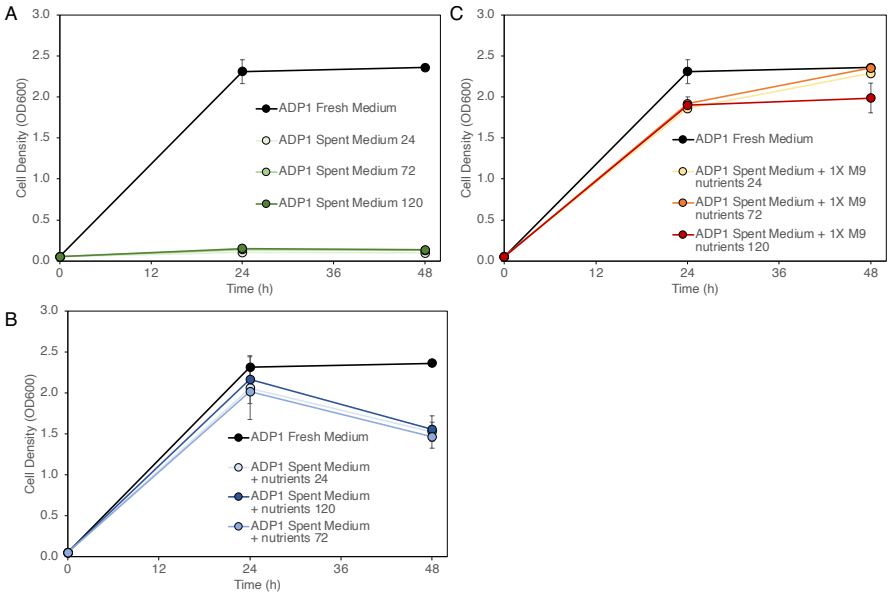

Figure 3-2 Cell density of ADP1 grown on spent medium collected from serially resuspended cultures at 24, 72, and 120 hours.
Spent media samples were either (A) used as is, (B) supplemented with nutrients and carbon (20 mM POB and 10 mM glucose), or (C) supplemented with M9 salts and carbon. Fresh M9 medium contained 20 mM POB and 10 mM glucose. N = 4 biological replicates. Error bars are S.E.M.

3.3.3 Nitrogen availability limits ADP1 growth in carbon-rich M9 minimal medium

To identify the limiting nutrient(s), ADP1 was grown in medium for which a single nutrient was excluded per condition. The initial 24 hours of growth was conducted in identical M9 medium for all conditions. At 24 hours, cultures were given a bolus of nutrients at 50% concentration relative to M9 medium excluding a single nutrient per condition. All conditions grew similarly until 72 hours, at which point the negative control and ammonium chloride deficient conditions exhibited reduced cell density relative to the positive control and all other conditions [Fig. 3-3A]. This reduced culture density persisted through 96 hours. At 96 hours, all conditions were supplemented with a bolus of the withheld nutrient. Upon feeding of ammonium chloride at 96 hours the cell density of the

ammonium chloride deficient condition reached a similar OD to the positive control by 120 hours, indicating full growth recovery. The monopotassium phosphate deficient condition experienced a decrease in cell density upon addition of potassium phosphate at 96 hours and remained at a reduced cell density relative to the positive control for the duration of the cultivation [Fig. 3-3B]. All other culture conditions grew similarly to the positive control.

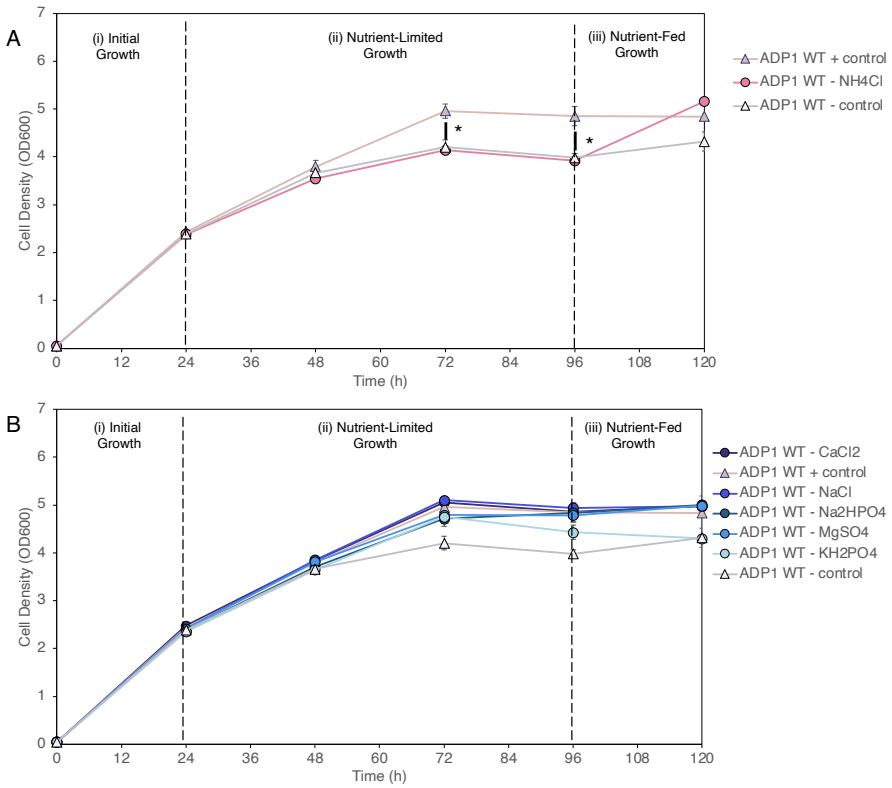

Figure 3-3 Growth and treatment conditions for nutrient elimination cultures.

Cell growth, shown by OD600, was tracked over 120 hours. Growth from 0 - 24 hours occurred in identical media for all conditions. Growth from 24 – 96 hours occurred in media supplemented with additional amounts of all nutrients except for one eliminated nutrient per condition. Growth from 96 – 144 hours occurred in media supplemented with the eliminated nutrient from each

condition. (A) Condition with excluded ammonium chloride. (B) Conditions with excluded buffering salts and micronutrients. N = 4 biological replicates. Error bars are S.E.M.

3.3.4 Bioreactors with improved aeration and nutrient feeding increases ADP1 culture density and growth rate

To investigate the impact of improved process control, ADP1 was cultured in 1L bioreactors at a working volume of 0.5L. The first 24 hours of cultivation were conducted in batch mode to utilize initial carbon before switching to fed-batch mode with feeding of 0.25M POB. In modified bioreactor M9 minimal medium with continual substrate drip-feed comprised of solely POB, ADP1 reaches OD 4.1 ± 0.3 at 72 hours [Fig. 3-4A]. Nitrogen is depleted by 48 hours, and carbon assimilation continues through 96 hours, at which point POB began accumulating [Fig. 3-4A-B].

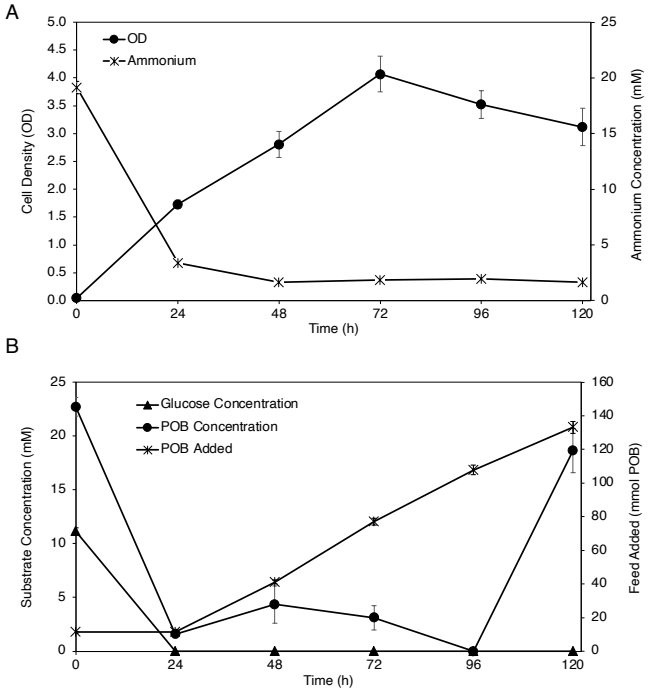

Figure 3-4 ADP1 WT growth and substrate assimilation in bioreactor cultures fed with solely 0.25M POB without supplementation of additional nutrients.

(A) Growth as shown by OD600 and nitrogen depletion as shown by ammonium concentration in non-nutrient fed bioreactor cultures. (B) Substrate concentrations for POB and glucose. N = 3 biological replicates. Error bars are S.E.M.

To evaluate the impact of supplying additional nutrients, particularly nitrogen in the form of ammonium chloride, ADP1 was cultured with supplementation of nutrients as well as POB. When cultures were provided with additional nutrients, including nitrogen, magnesium, calcium, and potassium, as a bolus alongside drip-fed POB, growth is prolonged and overall cell density is increased to 11.3 ± 1.6 by 360h of cultivation [Fig. 3-5A]. Ammonium chloride was supplemented at a higher concentration than other nutrients to avoid nitrogen limitation without overfeeding other

nutrients. The concentration of POB feed was also increased from 0.25M to 0.5M to reduce the dilution rate. POB was assimilated steadily and maintained at a concentration below 10 mM throughout cultivation [Fig. 3-5B]. However, in a low agitation setting (250 rpm), oxygen becomes limiting and slows growth such that cell density increases at a linear rate of 0.04 ± 0.00 OD/h. The linear growth is due to limited oxygen availability resulting from low mass transfer of oxygen, which is not dependent on cell density.

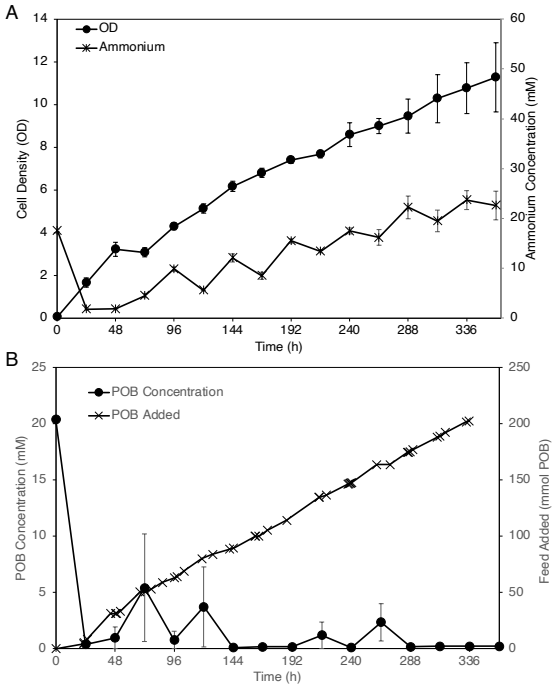

Figure 3-5 ADP1 WT growth and substrate assimilation in bioreactor cultures fed with 0.5M POB with supplementation of additional nutrients.

(A) Growth as shown by OD600 and nitrogen concentration as shown by ammonium concentration in nutrient fed bioreactor cultures. (B) Substrate concentrations for POB and total amount POB. N = 2 biological replicates. Error bars are S.E.M.

To address poor oxygenation, we implemented cascade agitation from 250 – 1200 rpm to maintain a dissolved oxygen concentration of 20% and increased the flow of sparged air from 1 VVM to 3 VVM. Carbon (0.5M POB) was fed at a rate that did not reduce oxygen below 20%. Using this strategy, ADP1 grew 13-fold faster at a linear rate with carbon limitation of 0.53 ± 0.02 OD/h, and cultures reached a cellular density of OD 6.8 ± 0.5 by 34 hours, or within 10 hours of beginning the fed-batch phase [Fig 3-6A]. Ammonium chloride and other nutrients were added as boli as previously described at 24 h to avoid nutrient limitation. To evaluate the impact of increasing agitation rate on ADP1 cell viability, ADP1 cells were plated after being subjected to 600, 800, 1000, and 1200 rpm agitation rates for 20 minutes prior to culture sampling. Colony forming unit counts for each condition did not differ significantly [Suppl. Fig. B-1]. However, at 1200 rpm, the rate of foam formation increased significantly, so moving forward a maximum agitation rate of 1000 rpm was used.

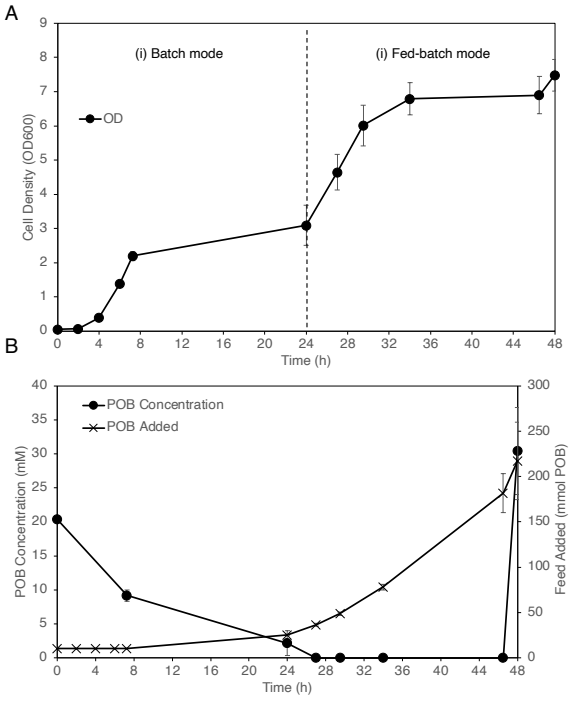

Figure 3-6 ADP1 WT growth and substrate assimilation in bioreactor cultures fed with 0.5M POB with supplementation of additional nutrients and cascade agitation to maintain dO_2 at 20%.
(A) Growth as shown by OD600 and nitrogen concentration as shown by ammonium concentration in nutrient fed bioreactor cultures. (B) Substrate concentrations for POB and total amount POB. N = 2 biological replicates. Error bars are S.E.M.

To further streamline ammonium chloride addition and to maintain a stable concentration of ammonium chloride in the medium, we next implemented co-feeding of POB and ammonium chloride. The yield of ADP1 cells from carbon and nitrogen were calculated from the previous 0.25M POB cultivation as 97.8 mM carbon/OD and 3.9 mM nitrogen/OD. Based on the ratio of the carbon to nitrogen yields, for 0.72M POB, the ammonium chloride concentration was 0.2M. Using this feeding strategy and adjusting agitation up to 1000 rpm to maintain dO2 at or above 20%, ADP1 grew to a final OD of 15.8 +/- 0.4 by 23 hours [Fig. 3-7A]. After reaching the maximum

agitation rate of 1000 rpm, the culture OD increased linearly by 0.88 ± 0.06 OD/h, representing a further 1.6-fold increase in ADP1 growth from the nitrogen bolus-feeding strategy. Throughout cultivation, POB was maintained below 10 mM with steady consumption of POB through 23 hours [Fig. 3-7B]. Ammonium was consumed at a rate faster than it was fed, necessitating the supplementation of additional ammonium chloride via bolus at 8.4 and 18.6h [Fig. 3-7C], thus indicating the concentration of ammonium chloride in the drip-fed solution should be increased. Based on these data, the yields of ADP1 cell biomass from nitrogen was recalculated and found to be 4.6 mM nitrogen/OD. This yield coefficient is also consistent with the initial baseline data from non-nutrient fed ADP1, for which the yield was 4.7 mM nitrogen/OD. By implementing co-feeding of ammonium chloride and POB alongside maintaining dO2 at or above 20% with higher agitation rates, we grew ADP1 to OD 15.8 +/- 0.4 within 23 hours, which is the maximum reported culture density for ADP1 in any growth format.

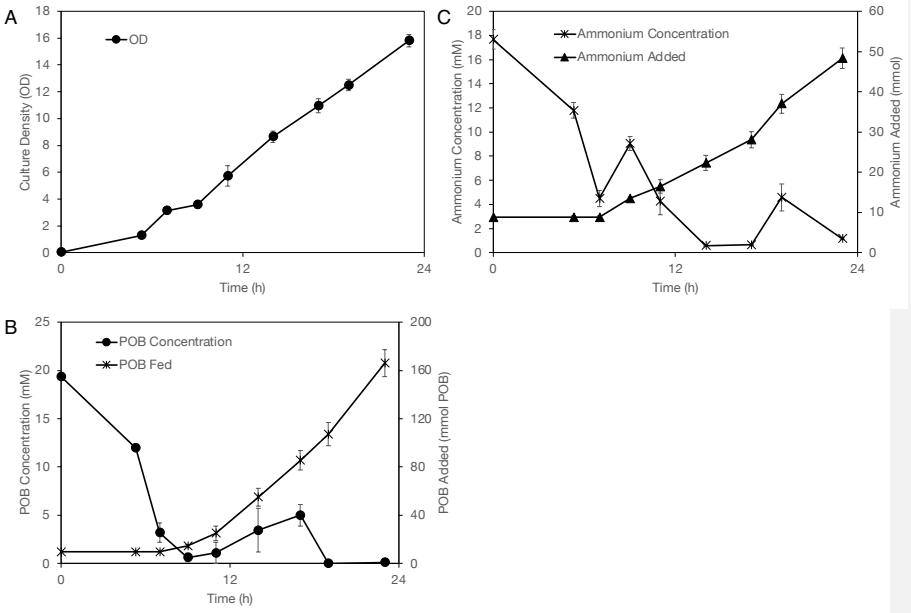

Figure 3-7 ADP1 WT growth and substrate consumption in bioreactor cultures fed with 0.72M POB with supplementation of additional nutrients and agitation between 250 – 1000 rpm to maintain dO_2 at 20%.

(A) Growth as shown by OD600. (B) Substrate concentrations for POB and total amount POB. (C) Nitrogen concentration and nitrogen addition as shown by ammonium concentration and ammonium addition in nutrient fed bioreactor cultures. N = 2 biological replicates. Error bars are S.E.M.

3.4 Discussion and Conclusions

In this study, we present strategies to address critical needs in scale up of ADP1 by (i) evaluating ADP1 for the presence of growth inhibition due to nutrient limitation or small molecule inhibition, (ii) identifying nutritional requirements for high density cell growth, and (iii) developing strategies for nutrient feeding and aeration to support rapid growth and high cell densities in bench-scale bioreactors without costly oxygen supplementation or specialized media.

Our findings indicate that ADP1 growth is limited primarily by nutrient depletion, and that nitrogen is the most limiting nutrient in M9 minimal medium. We hypothesized that different modes of growth inhibition would result in characteristic growth traits [Fig. 3-1A]: density-dependent inhibition could result in gradually lower density gains per cycle; decreasing cell density could indicate cytotoxic effects; nutrient limitation would result in consistent density gains each cycle. The observed growth characteristics most closely resemble the nutrient limited growth mode, with cell density increasing by a similar amount each cycle and no consistent reductions in density during each cycle. Though quorum sensing mechanisms have been documented in *Acinetobacter* species [86–88], they do not appear to impact growth under these conditions. Furthermore, growth inhibition by accumulation of inhibitory small molecules was not observed for the tested conditions. Variability in culture density and variability in density gains per resuspension-growth cycle increased with increasing cell density. This was potentially due to increasing difficulty in pelleting cultures, resulting in longer centrifugation times and poorer recovery of cells from culture

broth. However, due to the lower overall growth at higher cell densities, further evaluation was required to determine if inhibitory compounds effect cell growth.

By growing ADP1 in spent medium with additional nutrients, we ruled out the presence of density-dependent small-molecule based inhibition. Growth in spent medium with added nutrients and buffering salts was similar to growth in fresh medium, while growth in spent medium without added nutrients was minimal, supporting the hypothesis that the primary cause of growth limitation in spent medium was nutrient depletion. Furthermore, buffering capacity appeared to be a factor in growth for spent medium cultures [Suppl. Table B-1], and inclusion of 1X M9 buffering salts in spent medium maintained culture density from 24 – 48 hours, while cultures with solely additional micronutrients (magnesium sulfate and calcium chloride) and nitrogen (ammonium chloride) experienced reduced culture density in the same timeframe. All cultures assimilated carbon in similar amounts. Thus, it is likely that the spent supernatant did not contain any notable growth inhibitors, and that inclusion of nutrients and buffering salts found in M9 salts may be beneficial for sustaining ADP1 growth.

We next identified the primary limiting nutrient as nitrogen. Identification of the growth-limiting nutrient(s) is vital to enable targeted nutrient supplementation, thus avoiding unnecessary costs and reducing osmotic stress associated with accumulation of non-limiting nutrients. During growth on nutrient-restricted medium, a reduced culture density was observed by 72 hours for the ammonium chloride-deficient condition as well as the negative control compared to all other nutrient-restricted conditions and the non-restricted positive control. At 96 hours, the cultures were provided with a bolus of solely the eliminated nutrient. We postulated that if the supplemented nutrient is the sole factor limiting growth, supplementation of it would enable growth recovery. The ammonium chloride condition recovered fully with the addition of ammonium chloride, supporting our hypothesis that nitrogen had been depleted and thus limited ADP1's culture density. At the

observed cell densities, no other nutrients appeared to be limiting, though with higher culture densities, supplementation of additional nutrients may be required to sustain metabolic activity.

We demonstrated that targeted nutrient supplementation increases culture densities in bioreactor cultivation. Initially, we conducted bioreactor growth in a fed-batch mode with a feed comprised of solely pH-adjusted POB to determine the cell yield on initial nutrients before cultivating ADP1 with feeding of POB and dilute nutrients. With nutrient feeding, the growth phase was prolonged from 72 hours to 360 hours with steady linear growth. Culture density was also increased 2.8-fold from OD 4.1 to 11.3, which supports our hypothesis that the targeted nutrient strategy is sufficient to maintain non-limiting nutrient levels.

By increasing the rate of sparging from 1 VVM to 3 VVM and by implementing cascade agitation from 250 – 1200 RPM, we increased the linear growth rate of ADP1 significantly without the use of pure oxygen or costly media supplements. Inadequate oxygen transfer limits the growth rate of ADP1 in both shake flask and bioreactor cultivation. Industrial processes demand temporal efficiency alongside economical resource allocation. Therefore, increasing the growth rate by altering process parameters, namely sparging rate and agitation rate, is critical to establishing ADP1 as an attractive industrial chassis. The aeration strategies in this study allowed ADP1 to reach OD 6.8 in 34 hours compared to OD 11.3 in 360h hours with low sparging and agitation. The fastest observed linear growth was at a rate of 0.53 OD/h, relative to the growth at low oxygenation of 0.04 OD/h representing a 13-fold increase in linear growth rate. Though it requires higher gas flow rates and increased power usage for agitation, this strategy avoids the demand for pure oxygen supplementation, which drastically reduces process costs compared to processes with supplemental pure oxygen. Furthermore, we showed that agitation rates can be increased up to 1200 rpm without impacting cell viability [Suppl. Fig. B-1], which has not been previously explored when applying cascade agitation to ADP1 [33]. These findings demonstrate the advantage of ADP1's cellular robustness when applied to industrial processes.

Lastly, we combined the feeding of ammonium chloride and POB to further improve growth and to establish stable carbon and nitrogen concentrations in the medium. Streamlining of nutrient and carbon supplementation is desirable for industrialization as it enables process automation. Nitrogen availability plays a key role in directing carbon flux in ADP1, particularly for acetyl-CoA derived products[89], thus maintaining and tuning nitrogen availability is crucial to improving product yields. Furthermore, coupling carbon to nitrogen feeding enables tuning nitrogen supplementation to ADP1's carbon consumption, upon which it directly depends. Here, the nitrogen to carbon ratio in feed was determined based on biomass yields on each nutrient during ADP1 growth on POB as the sole carbon source. By introducing a constant drip feed of POB and ammonium chloride at 0.72M and 0.2M, respectively, in tandem with previous strategies for improved aeration we achieved an OD of 15.8 by 23 hours of growth with a maximum linear growth rate of 0.88 OD/h. This is the highest culture density for ADP1 reported to date in any growth format.

Further improvements may be achievable by introducing a constant drip feed of other essential nutrients, as, for these cultivations, bolus feeding of other M9 nutrients including potassium phosphate, sodium phosphate, iron, magnesium chloride, and calcium chloride were included to ensure no other nutrients became limiting. Future directions should explore ADP1 cell yields on each of these nutrients to establish bounds for automated feeding. Lastly, the use of POB as a substrate offers the advantage of being a simple single-compound carbon source, but its applications for industrial processes are limited by both solubility of POB, which is 0.72M, and by the purity of available lignin-derived waste streams. Future work should evaluate cell yields not only on a singular carbon source, but on mixed feedstocks more closely resembling actual lignocellulosic waste like mock APL[90].

Here we describe a scale up strategy to grow rapidly ADP1 to high density in inexpensive minimal medium on lignin-derived aromatic carbon. We first evaluated ADP1 for density- and/or non-medium-dependent growth inhibition and found that ADP1's growth patterns support the

hypothesis that ADP1 growth in M9 minimal medium is nutrient limited. This hypothesis was further supported by the lack of growth inhibition for ADP1 grown in spent medium with supplemented nutrients and buffering salts. We next identified nitrogen as the most limiting nutrient in M9 medium. Moving to bench-scale 1L bioreactors, we evaluated the impact of various culture parameters including agitation, sparging rate, and nutrient supplementation on ADP1 growth. By increasing the sparging rate of air and adjusting the agitation rate between 250 and 1000 rpm, we enabled maintenance of dO_2 at or above 20% without the use of costly oxygen supplementation. Further, by introducing co-feeding of nitrogen alongside the aromatic carbon source, POB, we attained a culture density of 15.8 and a linear growth rate of 0.88 OD/h, representing a 3.9-fold and a 22-fold improvement in final OD and linear growth rate, respectively.

This work represents a crucial step towards scale up and industrialization of ADP1. ADP1 represents a wellspring of potential for upgrading lignocellulosic feedstocks to value added products. However, as an obligate aerobe and a non-model organism, scale-up is a significant obstacle to establishing ADP1 as an industrial chassis strain. By demonstrating simple and generalizable methods to increase culture density and implement strategic nutrient feeding, this study will enable future efforts to bring industrialization efforts to fruition.

4. A model-guided metabolomics workflow improves metabolic annotations in *Acinetobacter baylyi* sp. ADP1

Author's Note: *This work is currently in preparation for publication. Dr. Jon Strutz will be the 1st author and performed all computational work. I will be the 2nd author and performed the final experimental validation, which is currently in progress.*

4.1 Introduction

Much of microbial metabolism is currently unknown, also termed "underground metabolism" [91]. Unearthing this underground metabolism could lead to new, exciting opportunities for bioproduction processes including identification of novel biomarkers [92], production of novel compounds [93], and engineering of newly discovered pathways and associated enzymes [94]. Thus, the process of cataloguing the metabolites in a given organism, metabolite annotation, is of critical importance [95]. However, metabolite annotation is challenging, as novel metabolites will not typically be present in metabolomics databases (e.g. MassBank), and the space of all possible chemical structures is vast [96,97].

In order to reduce this space of possible structures, untargeted metabolomics techniques are often used. Usually, a chromatography step is first used to separate compounds in a biological sample on the basis of charge, polarity, and other interactions. Then the mass-to-charge ratio, or m/z, is typically measured by mass spectrometry. Sometimes, compounds then go through a round of fragmentation where the spectra of a compound's fragments are measured, resulting in MS2 spectra. The raw metabolomics data is processed, resulting in a final set of peaks, each with a measured m/z, retention time, and sometimes MS2 spectra. Metabolite annotation is the process of annotating each of these peaks with chemical identities.

An accurate m/z value often reveals the likely molecular formula but not the structure. Thus, retention time and MS2 are used to aid in structural identification. While studies measuring liquid

chromatography retention times for known compounds are numerous, retention times are heavily influenced by chromatographic conditions and the column used so are difficult to compare across studies (although recent work has helped correct for these factors) [98–102]. Moreover, chromatographic and spectral databases do not typically contain retention times or spectra for novel metabolites. Additionally, because biological databases such as KEGG and MetaCyc contain only known biological compounds, more generalized chemical databases such as PubChem are often used when searching via m/z alone [103–105]**.** However, most PubChem molecules have low natural product (NP) likeness scores, indicating non-biological origin, making it unsuitable as a database of biological candidate structures [106,107].

To address this problem, Metabolic *In silico* Network Expansions (MINEs) were developed in 2015 and have recently been improved [108]. MINEs contain candidate structures that are predicted by applying common biochemical transformations to all known metabolites. Thus, while not every predicted compound will exist within metabolism, each one is derived using rules based on known biochemistry, making MINEs a more suitable database for annotating unknown peaks in metabolomics datasets, as shown in [108]. Other techniques such as Biotransformer and RetroRules can predict novel products as well, although Biotransformer is better suited for more targeted applications (e.g. prediction of xenobiotics) and while RetroRules is quite comprehensive, an even more comprehensive ruleset was recently developed for MINE [109–111]. Finally, Menikarachchi et al. created the IIMDB database of predicted enzyme products, but it primarily focuses on predicting unknown features of human metabolism so may not be suitable for metabolomics datasets from bacterial samples [112]. The MINE databases are accessible via a web app, and each MINE compound is linked with MS2 spectra, predicted *in silico* with CFM-ID 4.0, allowing for MS2 based searches as well as searches by m/z alone [113].

While MINE 2.0 utilized a more comprehensive set of reaction rules than MINE 1.0 during MINE construction, there were often still too many candidates for a given peak. While use of MS2

spectra was reported to significantly reduce this issue, MS2 spectra are not always measured for every peak in a dataset or may be of low quality [114–116]. In this work, we took an organism- and dataset-specific approach to develop a pipeline to generate a more targeted and credible set of candidate compounds for a dataset without the use of MS2 spectra (Figure 4.1).

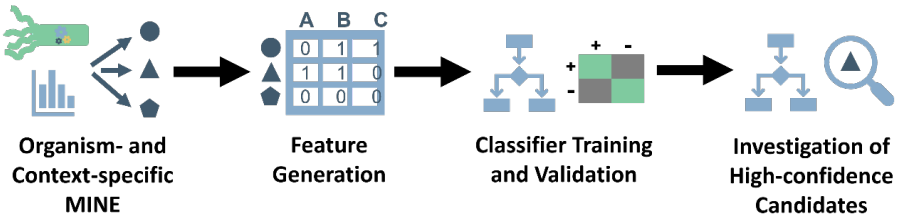

Figure 4-1 An organism- and context-specific machine learning approach for metabolite annotation

Compounds are first predicted from those that exist in a given organism using a ruleset derived from that organism's genes and reactions. This predicted compound set is filtered to those that map to peaks (based on m/z) in a metabolomics dataset for that organism. A machine learning classifier is used to reduce the size of the candidate set and is trained on features calculated for candidates associated with known (experimentally validated) peaks from the dataset. Calculated features include non-metabolomics quantities such as features derived from metabolic modeling and reaction feasibility prediction. This trained classifier is then used to predict the most likely candidates for unknown peaks in the dataset. These high-confidence candidates are then targets for further investigation and testing experimentally using a more targeted metabolomics approach (e.g. with a reference standard, if available).

We first curated a set of organism-specific reaction rules relevant to *Acinetobacter baylyi* sp. ADP1 (hereafter simply referred to as ADP1), a highly engineerable soil microbe with aromatic degradation pathways and thus potential as a platform strain for lignin valorization [29,68,117]. While organism-specific reaction rule generation based on the organism's reactome has been reported elsewhere, we took a more comprehensive approach by considering both the ADP1 reaction network and the ADP1 genome [118]. We then applied this set of ADP1-specific reaction rules to known metabolites within ADP1 to create a MINE database containing candidate compounds for a literature metabolomics dataset for ADP1 [38]. We combined this database of predicted reactions

with the ADP1 genome-scale model (GEM) to create an Extended Metabolic Model (EMM), as done in Hassanpour *et al.* [118]. However, we were interested not only in using the EMM as our candidate set of structures (as done in Hassanpour *et al.*) but also testing whether the EMM could be used itself to predict features of each candidate (such as maximum theoretical yield) that may be useful for metabolite annotation. Thus, after designing and calculating these features (and others), we fed them into a random forest classifier which was trained to filter out unlikely candidates and was validated against experimentally validated peaks in the ADP1 dataset. This work highlights the importance of non-metabolomics features for metabolite annotation, such as features generated by metabolic modeling. Finally, after applying the trained classifier to unknown peaks in the ADP1 dataset, we filtered the predicted set of 4,697 candidates down to just 96 high-confidence candidates. We are currently attempting to experimentally validate some of these candidates against reference standards via LC-MS/MS.

4.2 Materials and Methods

4.2.1 Acinetobacter baylyi sp. ADP1 Metabolic Model Preprocessing

The genome-scale metabolic model (GEM) for *Acinetobacter baylyi* sp. ADP1, titled "iAbaylyiv4", was downloaded from the BioModels database and cleaned up to comply with current (L3V1) SBML standards (see Appendix C.1 for details) [119,120].

4.2.2 Acinetobacter baylyi sp. ADP1 Metabolomics Dataset

The untargeted metabolomics dataset used in this work was taken from a previously published study where ADP1 was grown on quinate and succinate [38]. This dataset contained 448 validated peaks, 102 of which had been experimentally identified. While MS2 data was used to aid identification in the original work, the MS2 data was not stored online with the publication. Upon our request of the MS2 spectra, the authors provided the raw metabolomics data. MS2 spectra were obtained in the original study by processing this raw data using XCMS Online [38,121].

Attempts were made to reproduce this workflow using XCMS Online (as well as other tools such as MZmine), but we were not able to reproduce the reported MS1 peak list, let alone obtain MS2 spectra for all reported MS2 peaks [121,122]. Because of these issues, for this work we used only the MS1 data provided with the original publication.

4.2.3 Organism-Specific Reaction Rule Generation

A combined reaction- and gene-based approach was used to filter down a comprehensive set of reaction rules down to just those that are most relevant for *Acinetobacter baylyi* sp. ADP1. Recently, JN1224min, a comprehensive biochemical reaction ruleset derived from reactions in MetaCyc was published [104,110]. This ruleset was used to generate a larger but more specific (less promiscuous) ruleset by creating new rules for each rule in JN1224min where each new rule requires specific spectator atoms (atoms not participating in any bond changes but thought to be required for the reaction to occur). This resulted in a set of 7,387 rules with high specificity. This high specificity allowed us to better determine which rules are most relevant for *Acinetobacter baylyi* sp. ADP1. First, every rule was compared against every "mappable" reaction in the ADP1 GEM. "unmappable" reactions were defined as those that were (1) transport reactions, (2) comprised of cofactors only, (3) composite reactions with 8 or more total substrates and products, or (4) reactions with substrates/products not able to be annotated with SMILES strings (common for larger molecules such as ACP complexes, tRNAs, etc.). If the bond changes, spectator atoms, and cofactors of a rule matched those of any "mappable" reaction from the ADP1 GEM, that rule was considered to be "reaction-mapped". RDKit, a python library, was used to carry out these comparisons [123].

In addition to this set of "reaction-mapped" rules, a set of "gene-mapped" rules was generated. First, a FASTA file was generated for each rule, containing the enzyme amino acid sequences associated with that rule. Of the 7,387 rules, 2,776 contained no associated enzyme

sequences and were not included as candidates for "gene-mapping". Specifically, MUSCLE was used to perform a multiple sequence alignment on the amino acid sequences of each rule [124]. Then, *hmmbuild* was used to build a HMMER profile for each rule, and *hmmsearch* was used to query that rule's profile against the 3,307 enzyme-coding genes in ADP1 to search for homology [125]. For each reaction rule, this produced a table of hits with E-values (expectation values) less than 10^{-1} (the default threshold). All of these tables were merged into a single, large table locally using a custom python script, where each row contained the rule name, the name of the ADP1 gene hit by that rule, as well as the E-value, score, and description of the gene. Multiple genes could be hit by the same rule (and multiple rules could hit the same gene). Rules with an E-value less than 10^{-35} for at least one hit were kept in the final "gene-mapped" set. See Appendix C.2, Figure C.1, and Figure C.2 for details on how this E-value threshold was chosen.

Finally, an additional set of 96 spontaneous reaction rules (which are organism-independent as they are derived from spontaneous biological reactions that do not require enzymes) were added to this set of ADP1-specific rules to create the final ruleset [126].

4.2.4 Metabolic In silico Network Expansion (MINE)

Pickaxe, our in-house reaction prediction software, was used to predict reactions by applying the final ADP1-specific ruleset to all compounds in the ADP1 GEM that contained SMILES strings. This created a Metabolic *In silico* Network Expansion (MINE) specific to ADP1. This MINE expansion was generated by running Pickaxe for 1 generation on 12 threads on a local machine (Windows Surfacebook 2, Intel i7-8650U CPU, 1.90 GHz, 16 GB RAM). The Pickaxe metabolomics filter was used after the expansion to filter the predicted compound set to only those that match at least one peak m/z in the ADP1 MS1 peak list. All precursor ion/adduct combinations were considered for each predicted compound, using the adducts reported in the publication associated with this dataset ([M+H]+, [M+NH_4]+, [M+Na]+, [M+K]+ for positive ionization mode

and [M-H]- for negative ionization mode) [38]. An m/z tolerance of 2 mDa was used, roughly equal to the tolerance used in the original publication (10 ppm) given a theoretical mass of 200 Da [38].

4.2.5 ADP1 Extended Metabolic Model (EMM) Construction

The set of metabolomics-filtered predicted compounds and their associated predicted reactions were added to the ADP1 GEM to create an extended metabolic model (EMM) [127]. Because Pickaxe does not consider the compartment of substrate metabolites during reaction prediction, only predicted reactions with substrates from the same compartment in the model were added (products were assumed to be in the same compartment as reactants). Boundary reactions allowing all predicted compounds to leave but not enter the system were also added to facilitate calculations like maximum theoretical yield of a candidate. In all, the model contained 8,301 metabolites and 11,000 reactions, 316 of which were gapfilling reactions (new reactions containing known ADP1 compounds as both substrate and product).

4.2.6 Feature Generation

In order to train a machine learning classifier to distinguish true metabolite annotations from false metabolite annotations for each peak, each compound-peak pair needed to be featurized. Three sets of features were designed and calculated for each candidate: features based on (1) metabolic modeling, (2) predicted reaction feasibility, and (3) metabolomics and other properties (Figure C.3). Metabolic modeling-based features were calculated using the ADP1 EMM. Modeling features include maximum theoretical yield (MTY) of a candidate compound, the number of experimentally validated compounds involved in flux-carrying reactions in the flux distribution predicted when calculating MTY, and the distance between the candidate and feed compounds in the metabolic network. Feasibility-based features were calculated using deepRFC, a tool to predict the feasibility of biochemical reactions [128]. The number of feasible reactions producing a candidate compound, the fraction of all reactions producing a candidate compound

that were classified as feasible, and whether any reactions producing a candidate were classified as feasible were all included as feasibility-based features. Metabolomics features include those based on predicted retention time as well as differences in peak intensity in the two feed conditions used in the metabolomics experiment for the ADP1 dataset. Other features include the number of unique precursor compounds for a candidate as well as molecular properties such as aromaticity. A full description of all calculated features can be found in Table C.1.

Some features were not able to be calculated for every compound-peak pair. In these cases, a value was either set to a default value (e.g., 0, 1) or sampled (with a preset random seed) from the other values calculated for that feature across all compound-peak pairs. This choice depended on the feature (see Appendix C.3 for justification of various imputation methods) [129]. The primary cause of inability to calculate feature values was when calculating metabolic model-based features for large secondary metabolites not connected to quinate or succinate in the metabolic model which affected roughly 20% of candidate-peak pairs (e.g., we could not calculate the metabolic network distance from quinate to a candidate derived from one of these secondary metabolites). All other incomplete feature columns had less than 5% candidate-peak pairs that had to be sampled or set to default values (Appendix C.3). Features were iteratively added, changed, and removed during model training and validation to improve classifier performance – the features described in this work are the final set of features used.

4.2.7 Random Forest Classifier – Training and Validation

The python library, sklearn, was used to train a random forest classifier on a dataset of known and predicted candidate compounds for 75 known peaks in the ADP1 dataset (not all 102 known peaks were associated with compounds that were in the model) [130]. This dataset consisted of 75 positives (experimentally validated compound-peak pairs) and 2,640 negatives (predicted compounds that happened to match an m/z within this set of known peaks). Because these

classes were unbalanced, during the bootstrapping step of tree generation, these two classes were weighted such that positives and negatives were equally likely to be sampled.

Model performance was quantified by the F-score, the harmonic mean of recall and precision, for out-of-bag samples. Out-of-bag samples were used as the validation set rather than holding out a separate validation set due to the low number of positive examples in the dataset (n=75) [131]. Similarly, there was no held-out final test set as it would likely not be sufficiently representative of the full set of positive examples. We decided that potentially biasing final performance metrics by not having a separate held-out (likely not sufficiently representative) test set was worth it as it allowed us to use all of the (very limited) dataset for model training.

Class threshold tuning and feature reduction were both used to improve model performance and interpretability [132,133]. The model was initially trained on all features with 200 trees, class balancing, and at one of 10 random states. The 10 different random states were used to account for potential variation in performance from bootstrapping. This resulted in 10 initial models with similar performance. To investigate the effects of feature reduction, each model was then retrained on the top *n* features only, where *n* was 2, 5, 8, 10, 15, or 26 (all features). The top features were determined by ranking features by their calculated Gini importance from the initially trained models. In addition, a PR (precision-recall) curve was generated for each *n* (for each of the 10 models) by varying the classification threshold from 0 to 1. These PR curves were used to determine the lowest number of features (i.e., the simplest model) that could be used while maintaining model performance. This was done while simultaneously choosing the best threshold for classification at each *n*. To quantify this, the best F-score for every PR curve (for each *n*) was calculated. The best performing set of 10 models (at 10 different random states) were found to be those retrained on the top 10 features with a threshold of 0.30. A final model was then trained at a different (arbitrary) random state (42) on the top 10 features with this threshold. This model's

out-of-bag results were used to calculate the final recall, precision, and F-score metrics as well as Gini feature importances for each of the 10 features.

4.2.8 Random Forest Classifier – Prediction on Unknowns

The optimized random forest classifier was then used to predict the class probabilities for all peak-compound pairs for unknown peaks in the ADP1 metabolomics dataset (those without an experimentally identified compound). Those with a probability above 0.30 were considered "hits". Only this small set of positively labeled compounds was considered for final analysis.

4.2.9 Analysis of High-Confidence Hits

In order to prioritize which hits to test experimentally, a literature and database search was performed for each hit. Specifically, any associated literature (e.g., studies annotated on the compound's PubChem page) was investigated. In addition, we looked at the predicted reactions producing each predicted compound – in particular, we searched for similar known reactions in databases like MetaCyc, KEGG, and BRENDA by looking at the EC (Enzyme Commission) numbers and genes associated with each reaction rule [104,134,135]. We also made use of homology-search tools like BLAST to check for homologous enzymes (e.g., those that might catalyze one of these similar known reactions) in the ADP1 genome [136]. The Protein Data Bank (PDB) was used as well to check for enzyme-ligand interactions [137]. Based on this research, hits with significant literature support, often from multiple sources, were prioritized for experimental validation.

4.2.10 Experimental Validation of High-Confidence Hits

4.2.10.1 Cloning Methodology

All primers and plasmids used in this study may be found in Appendix C Tables C-2 – C-3. Genomic homology regions were amplified from the ADP1 genome via colony PCR, with the DNA template being a colony of wild type ADP1 suspended in 50 μL of nuclease free water. All

PCRs were conducted with PrimeSTAR Max Master Mix (2X) (Takara Bio). PCR products were run via electrophoresis on a 1% agarose gel at 90V for 25 minutes and extracted using the GeneJet Gel Extraction kit (Thermo Scientific) to purify DNA.

Knock-outs were performed using genome homology-flanked selection cassettes, as described in Biggs *et al.* [29]. Genome homology upstream and downstream of the gene targeted for knock-out were amplified from the ADP1 genome alongside a selection marker (either kanamycin or chloramphenicol resistance). The parts for each selection cassette were purified via gel extraction and stitched together using overlap PCR. The selection cassettes were purified via gel extraction and transformed into ADP1 with a minimum of 1000 ng of DNA [29]. Transformants were screened via colony PCR, and potential hits were sequence verified.

ACIAD1826 (*pauA*), a pimeloyl-CoA synthetase is associated with synthesis of 3-hydroxyadipic acid. The knock-out cassette for ACIAD1826 was comprised of 500 basepairs of ADP1 genome homology upstream of ACIAD1826, amplified with primers ECA158/159, a kanamycin resistance gene from pBWB162, amplified with primers ECA160/161, and 500 basepairs of ADP1 genome homology downstream of ACIAD1826, amplified with primers ECA162/163.

ACIAD0381, a putative flavoprotein monooxygenase, is associated with synthesis of 3-hydroxyanthranillic acid. The knock-out cassette for ACIAD0381 was comprised of 500 basepairs of ADP1 genome homology upstream of ACIAD0381, amplified with primers ECA146/147, a kanamycin resistance gene from pBWB162, amplified with primers ECA148/149, and 500 basepairs of ADP1 genome homology downstream of ACIAD0381, amplified with primers ECA150/151.

ACIAD0984, a putative hydroxylase involved in salicylate metabolism, is associated with synthesis of 3-hydroxyanthranillic acid. The knock-out cassette for ACIAD0984 was comprised of 500 basepairs of ADP1 genome homology upstream of ACIAD0984, amplified with primers

ECA152/153, a kanamycin resistance gene from pECA19, amplified with primers ECA154/155, and 500 basepairs of ADP1 genome homology downstream of ACIAD0984, amplified with primers ECA156/157.

ACIAD3540, a putative flavoprotein monooxygenase acting on aromatic compound, is associated with synthesis of 3-hydroxyanthranillic acid. The knock-out cassette for ACIAD3540 was comprised of 500 basepairs of ADP1 genome homology upstream of ACIAD3540, amplified with primers ECA188/189, a kanamycin resistance gene from pBWB162, amplified with primers ECA190/191, and 500 basepairs of ADP1 genome homology downstream of ACIAD3540, amplified with primers ECA192/193.

4.2.10.2 Cultivation Methodology

Wild type *A. baylyi* ADP1 was used for cell culture studies. Precultures were initially cultivated in LB medium by inoculating 5 mL of LB in a 14 mL Falcon tube with a single colony and grown overnight at 30°C and 250 rpm at 45° rotation. They were then diluted 1:100 into the relevant minimal medium and grown for an addition 12-18 hours. To inoculate cultures, precultures were pelleted at 6800 x g and 4°C and resuspended to an OD of 5 with fresh minimal medium. Cultures were inoculated to an initial OD of 0.05 and grown at 30°C and 250 rpm unless otherwise noted.

Cultivation medium was Medium for *Acinetobacter* (MA) [31 mM Na2HPO4, 25mM KH2PO4, 18mM NH4Cl, 41 μM nitrilotriacetic acid, 2 mM MgSO4, 0.45 mM CaCl2, 3 μM FeCl3, 1 μM MnCl2, 1 μM ZnCl2, 0.3 μM (CrCl3, H3BO3, CoCl2, CuCl2, NiCl2, Na2MoO2, Na2SeO3)] supplemented with 25 mM quinate. Precultures were grown in 5 mL MA medium in 14 mL falcon tubes at 30°C and 250 rpm for 24 hours prior to harvesting.

4.2.10.3 Metabolome Extraction and Analytical Methodology

Optical density measurements were taken via Synergy H1 microplate reader (BioTek) and Flat Bottom Clear Non-sterile 96-well plates (Fisherbrand). Cultures were diluted to reach an absorbance reading between 0.2 – 0.7 using sterile medium for accurate measurement of OD_{600}. Metabolome preparation was adapted from Stuani *et al.* [38]. Cultures were inoculated in 25 mL MA medium at OD 0.05 and grown to OD 0.2 at 30°C and 250 rpm. Cells were then captured on a filter (cellulose acetate, 25 mm, 0.45 μm), placed cell-side-up onto minimal medium agarose, and grown at 30°C to OD 0.8 (approximately 5 hours). Cell metabolism was quenched by placing the filter cell-side-down into a 15 mL falcon tube containing 5 mL of -30°C 80% LCMS-grade acetonitrile 20% LCMS-grade methanol and incubated for 15 minutes at -30°C. Quenching liquid was freeze-thawed 6 times in liquid nitrogen and 65°C water, then centrifuged in a vacuum centrifuge to evaporate liquid. Dry samples were suspended in 500 μL purified water, centrifuged at 2000 x g at 4°C for 10 minutes to pellet cell debris. Supernatant was transferred to a microtube and vacuum centrifuged to evaporate liquid, then dry samples were stored at -20°C until resuspension and analysis.

Dry samples were resuspended in 100% LCMS-grade acetonitrile at various dilutions to determine the appropriate dilution ratio (volumes ranged from 1 mL to 250 μL). A 500 μL resuspension volume was used for final cultivations. Samples were stored at 4°C after resuspension and analyzed within 24 hours. Samples were analyzed on an Agilent 1200 series LC Bruker AmaZon-X Ion Mass Trap MS. Liquid chromatography was conducted using a method adapted from Stuani *et al.* [38] with mobile phase A being comprised or 90% LCMS-grade water and 10% 10 mM ammonium formate, pH adjusted to 10 with ammonium hydroxide. Mobile phase B was comprised of 90% LCMS-grade acetonitrile and 10% LCMS-grade water with 100 mM ammonium formate, pH adjusted to 10 with ammonium hydroxide. An Agilent HILIC-Z column

was used for chromatographic separations. Standards of 3-hydroxyanthranilic acid and 3-hydroxyadipic acid were used for calibration and identification standards.

Diluted samples were injected on a 1290 Infinity II UHPLC System (Agilent Technologies Inc., Santa Clara, California, USA) onto a Poroshell 120 HILIC-Z column (2.7 µm, 100 × 2.1 mm) (Agilent Technologies Inc., Santa Clara, California, USA) for hydrophilic interaction chromatography separation (HILIC) which was maintained at 30 °C with a constant flow rate at 0.400 ml/min, using a gradient of mobile phase A (water, 10 mM ammonium acetate) and mobile phase B (90% acetonitrile, 10 mM ammonium acetate). The gradient program was as follows: 0 – 1 min, 80 %B; 1 – 3 min, 80 – 68 %B; 3 – 4 mins, 68 – 40 %B; 4 - 7 min, hold 40 %B; 7 – 7.10 min, 80 %B; 7.10 - 12 min, hold 80 %B. "MS-Only", negative and positive ion mode acquisition was conducted on the submitted samples on an Agilent 6545 quadrupole time-of-flight mass spectrometer equipped with a JetStream ionization source. The source conditions were as follows: Gas Temperature, 300 °C; Drying Gas flow, 12 L/ min; Nebulizer, 45 psi; Sheath Gas Temperature, 350 °C; Sheath Gas Flow, 11 L/ min; VCap, 3000 V; Fragmentor, 100 V; Skimmer, 65 V; and Oct 1 RF, 750 V. The acquisition rate in MS-Only mode was 3 spectra/second, utilizing m/z 121.050873 and m/z 922.009798 as reference masses.

Commented [EA1]: LCMS method

4.3 Results

4.3.1 Organism-specific expansion generates a more targeted candidate set for metabolite annotation

In order to predict a targeted set of candidate structures that are not only biologically relevant but also relevant specifically to ADP1, we used a ruleset derived from the ADP1 metabolic network and genome to generate a total of 1,710 ADP1-specifc reaction rules (Figure 4.2a) [104,110,135]. We found that the reaction-mapped and gene-mapped rulesets had significant overlap (252 of 323 reaction-mapped rules). This result gave us confidence that rules that were

gene-mapped but not reaction-mapped may also be relevant for ADP1. These two sets of rules were combined with an additional set of 96 spontaneous reaction rules (which are organism-independent as they do not require enzymes), totaling 1,710 reaction rules in the final set.

Applying these ADP1-specific rules to a set of 522 starting compounds from ADP1 to create a MINE resulted in 200,283 predicted compounds. Filtering this set of compounds to only those that matched at least one peak in the metabolomics dataset resulted in 6,102 unique candidate compounds (Figure 4.2b). There were 3,264 candidates for known peaks and 4,697 for unknown peaks for a total of 7,961 candidate-peak pairs (many compounds matched multiple peaks). 97 of 102 known peaks and 201 of 346 unknown peaks in the dataset had at least one predicted candidate compound. To evaluate the efficacy of organism-specific reaction rules (and starting compounds) at reducing the candidate set, we also looked at how many candidates we would have predicted using other sets of starting compounds (specifically, compounds in KEGG) and rules (a comprehensive ruleset derived from MetaCyc). Using our organism-specific approach resulted in a 34-fold decrease in the number of candidate compounds compared to using these KEGG starting compounds and MetaCyc-based reaction rules (Figure 4.2b).

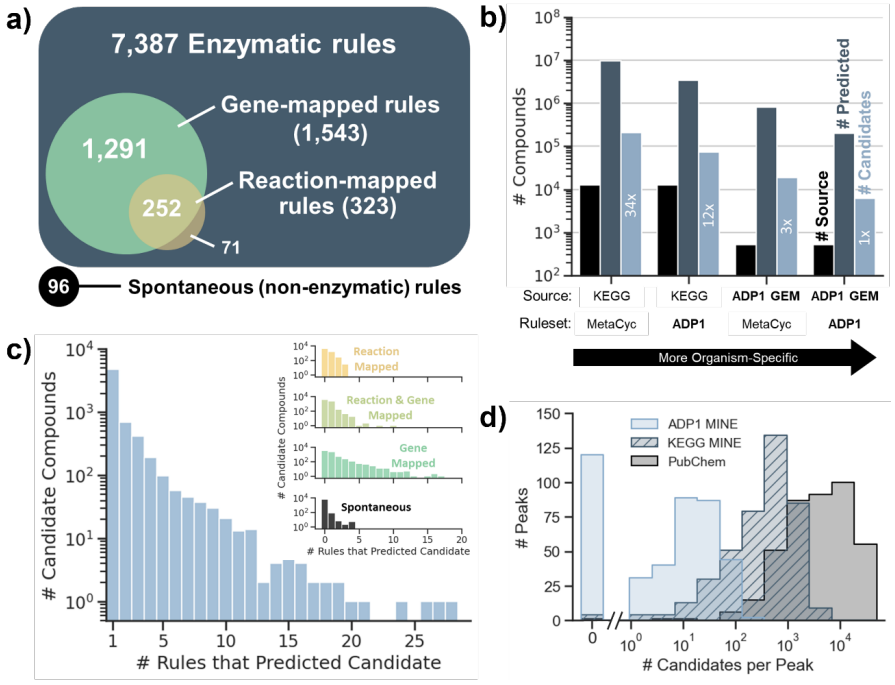

Figure 4-2 Organism-specific expansion generates a more targeted candidate set for metabolite annotation in *Acinetobacter baylyi* sp. ADP1

(a) A comprehensive ruleset is filtered down to just those rules that are most relevant for ADP1 (comprised of reaction-based, gene-based, and spontaneous rule subsets). The number of rules in each subset is shown in this Venn diagram. The final ruleset contains 1,710 rules. (b) This ruleset is then applied to compounds in ADP1 to predict novel reactions in ADP1. The total number of predicted compounds as well as the number of these that matched at least one peak in an ADP1 dataset (candidates) are plotted [38]. Expansions starting from either 522 ADP1 compounds or 12,688 KEGG compounds (using either the ADP1-specific ruleset or a ruleset derived from all of MetaCyc) are plotted as well to show the reduction in candidate set size by taking an organism-specific approach. (c) Many candidates in the ADP1-specific MINE are predicted via only a single reaction rule while others are predicted to be made via multiple (up to 28) predicted routes. More reactions are predicted by gene-based rules than other types (inset). (d) The ADP1 MINE contains a more targeted candidate set than the KEGG MINE (KEGG source with MetaCyc rules) or the PubChem database.

While most candidate compounds in the ADP1-specific MINE were only predicted to be produced via a single reaction rule, many were predicted via multiple routes (Figure 4.2c). Overall,

most candidate-producing reactions tended to be predicted via gene-mapped rules (Figure 4.2c, inset), with 34% of candidate compounds predicted using solely gene-mapped rules. However, reaction-based rules were utilized more often on an individual basis, with each reaction-mapped rule predicting an average of 15.9 candidate compounds (versus 5.5 for gene-mapped rules and 11.5 for rules that were both reaction- and gene-mapped). Although spontaneous rules were also present in the ruleset, they were rarely used (average of 1.3 candidates predicted per spontaneous rule).

Finally, we found that our targeted approach generated far fewer candidates per peak than less organism-specific approaches (Figure 4.2d). In particular, our most organism-specific approach (ADP1 MINE) resulted in an average of 13 candidates per peak. On the other hand, using PubChem resulted in an average of 459 candidates per peak. Additionally, many of the candidates predicted by PubChem are not biologically relevant. This has been quantified in previous work where it was found that the natural product scores of PubChem compounds are significantly lower than those of KEGG compounds, indicating non-biological origin [106,138]. Using more intermediate approaches like the KEGG MINE contains more biologically relevant compounds and performs better than PubChem, as reported elsewhere [106], but still suggests 164 candidates per peak, too many for annotation to be practical for many peaks in the dataset. While similar reductions in candidate space through the use of organism-specific, model-based approaches to reaction prediction have been reported elsewhere, the use of gene-mapped rules in this work significantly increases the number of unknown peaks (from 161 to 201) that are able to be covered [118]. While it should be noted that the ADP1 MINE did not generate any candidates for roughly a quarter of the peaks in the dataset, having a more practical number of candidates for those peaks that do have candidates is more desirable than having too many candidates for the vast majority of peaks as is the case for the KEGG MINE.

4.3.2 Machine learning workflow generates more targeted and credible candidate set for metabolite annotation

In order to further reduce the candidate set of the ADP1 MINE, a machine learning approach was used. A random forest classifier was trained on the candidates associated with 75 known peaks from the ADP1 metabolomics dataset. Each known peak's candidate set included both predicted compounds as well as that peak's experimentally validated compound (we ensured that this compound was able to be predicted from known precursors using the ADP1-specific ruleset).

Classifier performance was quantified by classifying out-of-bag samples, a form of cross-validation unique to the random forest algorithm. In order to optimize the classifier, feature reduction and threshold tuning were performed. Using the top 10 (of 26) features with a classification threshold of 0.30 maximized performance (F = 0.51) on out-of-bag samples across a range of 10 random seeds (Figure C.4). The top 10 features included features from all three feature sets (Figure C.5).

Figure 4.3a shows the confusion matrix for the out-of-bag validation of the final model. We found that the model filters out 96% of candidates (correctly) as true negatives (n=2,608) while retaining 48% of the experimentally validated compounds (n=36 of 75). Before applying the model, experimentally validated compounds made up less than 3% of the candidate set while after applying the model this proportion increased to 53%. Thus, this approach results in a more targeted, credible candidate set.

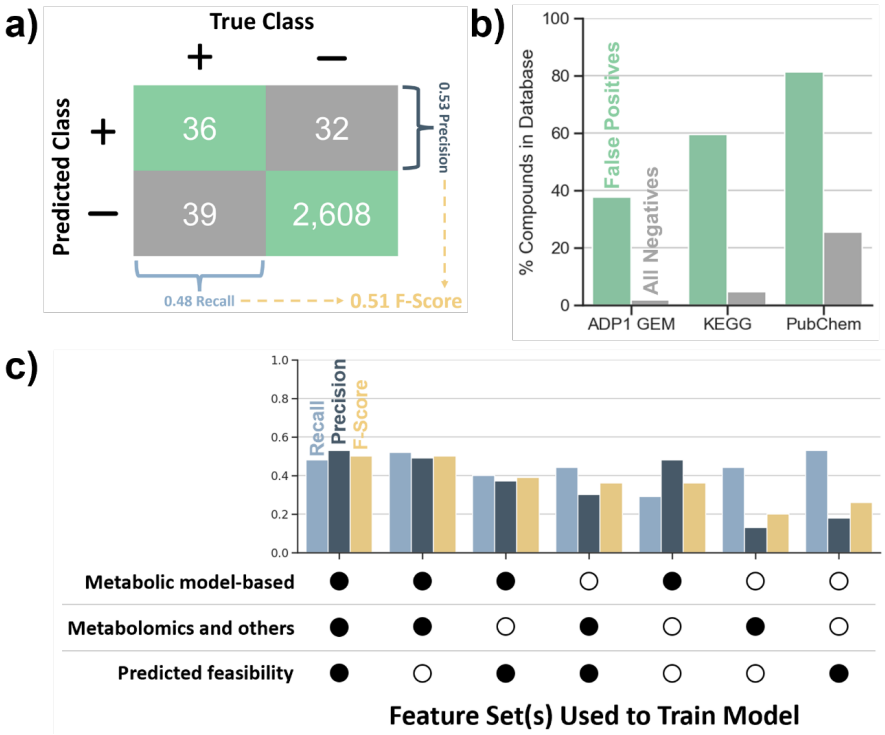

Figure 4-3 Machine learning workflow generates a more targeted and credible set of candidate structures

(a) The classifier, trained on all three feature sets, effectively filters out many true negatives while maintaining acceptable recall and precision. The confusion matrix displays the results of the random forest classifier on out-of-bag samples after feature reduction and classification threshold tuning. (b) Many false positives, while annotated as negatives in the dataset, actually do exist in ADP1 (but just were not measured or annotated with a peak in this specific ADP1 dataset) as well as in other databases like KEGG and PubChem. Negatives predicted as hits (false positives) are more likely to exist in these databases than negatives as a whole. (c) Incorporation of non-metabolomics features improves classifier performance. Performance (quantified by recall, precision, and F-score) was calculated from classification of out-of-bag samples after retraining the classifier on various combinations of feature sets, indicated along the x-axis (training included a feature reduction and threshold tuning optimization step).

There were also 32 predicted candidates that were classified as hits even though they were not experimentally validated for any of the known peaks in the dataset. We hypothesized that some of these false positives may actually exist in ADP1 but just happen to match one of the known peaks based on m/z. To test this, we calculated the percent of false positives that exist in ADP1 (specifically, whether it is in the genome-scale model) as well as the percent of all negatives (both false positives and true negatives) in ADP1. We found that while just 1.6% of all negatives were present in the ADP1 GEM, 38% of false positives were present (Figure 4.3b). This implies that although the model may predict a candidate compound that did not actually produce a given peak, that compound may still be present. It is also possible that multiple compounds (with the same m/z and retention times) produced a single peak, although this seems less likely. Analogous results were found when searching for false positives in larger non-organism-specific databases such as KEGG and PubChem (Figure 4.3b).

To gain insight into how the model was labelling candidates, we looked at the importances of the 10 features used by the final model (Figure C.5). Non-metabolomics features were found to be more important for prediction than metabolomics-based features. For example, the number of "feasible" (predicted via DeepRFC) reactions producing a candidate was found to be the most important feature [128]. Similarly, the total number of unique precursors for a predicted candidate was also important, albeit less so. Other metabolic network- and model-based features were also used in the final model. For example, the distance in the metabolic network between succinate (one of the feed compounds in the metabolomics experiment) and the candidate compound was the second most important feature overall. The metabolic distance from quinate (another feed compound) was important as well. Other model-based features were also useful for prediction: for example, the number of experimentally validated compounds that were predicted to have flux through them by the model (when optimizing for production of the candidate).

Because each peak was experimentally measured under one of two feed conditions (succinate or quinate), we also calculated features that capture any potential correlation between a candidate's distance from one of the feed compounds and the fold change (across feed compounds) for that peak. Thus, these features combined information from both the metabolic model and metabolomics dataset. Three of the ten top features fit into this category. Finally, one metabolomics feature was found to be important – the deviation between predicted and measured retention time for each candidate and peak, respectively (for details on retention time prediction, see Appendix C.5, Figure C.6, and Figure C.7) [139].

We also retrained (and reoptimized) the model on all possible combinations of the three feature sets to test their impact on model performance (Figure 4.3c). The highest F-score was obtained when using either all three feature sets or all but the feasibility-based features. However, slightly higher precision was obtained (at the expense of recall) when using all three feature sets. Importantly, while the model trained on all but the feasibility-based features performed well, the two other pairwise combinations of feature sets performed significantly worse.

We found that the metabolic model-based features made the largest impact on model-performance. Even when training the classifier on only the metabolic model-based features, an F-score of 0.36 was achieved, much higher than when training on only the metabolomics-based features (F = 0.20) or only the feasibility-based features (F = 0.26). That said, the model trained only on feasibility-based features did surprisingly well (F = 0.26) considering that there were only three calculated features in this set. Overall, these results show that information derived from sources other than the metabolomics dataset such as metabolic modeling and reaction feasibility prediction can be used to improve metabolite annotation workflows.

4.3.3 High-confidence predictions of metabolic features in *Acinetobacter baylyi* sp. ADP1

After training the classifier on known peaks, it was used to label candidate compounds that had been predicted for unknown peaks in the ADP1 metabolomics dataset. Of the 4,697 candidates for unknown peaks, 96 were classified as hits, thus further reducing our candidate set size by 98% (Figure 4.4a). Each compound in this small, high-confidence set was then investigated in detail. We looked at the predicted reactions, which reaction rules were used, homology between ADP1 enzymes and enzymes associated with each rule, as well as literature and vendor data. This process was time- and labor-intensive, and it would not have been practical for all 4,697 candidates, demonstrating the importance of developing and using tools able to filter out unlikely candidates.

This vetting process allowed us to narrow down this set of 96 candidates to the 5 that we found most compelling, warranting further investigation (Figure 4.4b). This set includes three compounds known to biology (but not ADP1): glutarate semialdehyde, 3-hydroxyanthranilate, and 2-ethylmalate; as well as two compounds unknown to biology (but present in PubChem): 3-hydroxyadipate and 4-hydroxy-2,5-dioxopentanoate. Information about each of these five candidate-peak pairs, evidence from the literature, hypotheses for candidate generation, as well as potentially relevant ADP1 genes are detailed below and also summarized in Table 4.1.

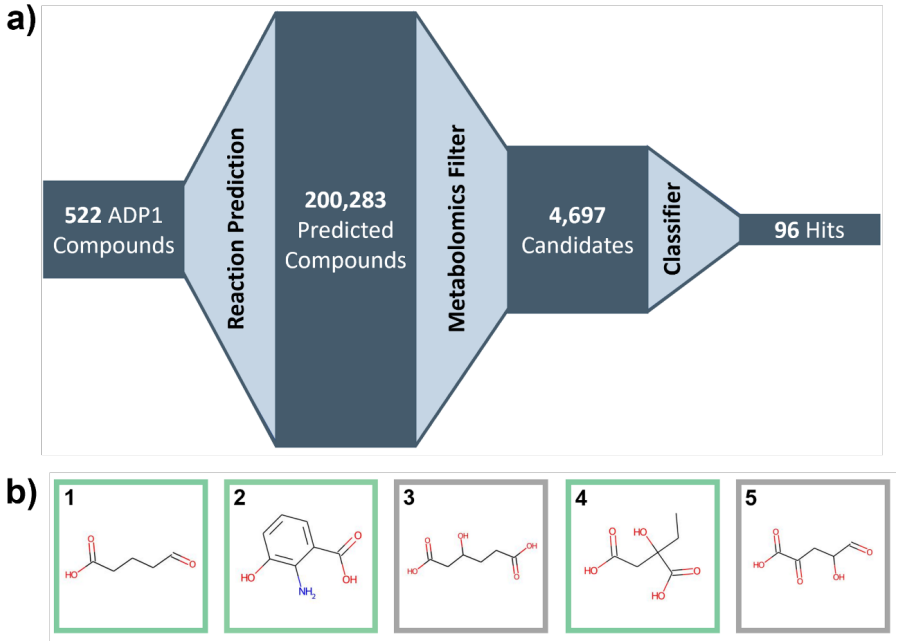

Figure 4-4 Classifier significantly reduces size of candidate set for unknown peaks, resulting in a small set of 96 high-confidence candidates, 5 of which were chosen for further investigation

(a) After predicting candidate compounds (n=3,801 across 202 peaks) for unknown peaks in the ADP1 metabolomics dataset, the classifier further reduced this set to a smaller set of the most likely candidates (n=96 across 49 peaks). (b) Of the 96 hits, 5 compounds were selected for further investigation. The predicted compounds highlighted in green are known to exist in other organisms but not in ADP1. Compounds highlighted in grey were not present in MetaCyc or KEGG but were present in PubChem. Compounds shown are (1) glutarate semialdehyde, (2) 3-hydroxyanthranilate, (3) 3-hydroxyadipate, (4) 2-ethylmalate, and (5) 4-hydroxy-2,5-dioxopentanoate. See Table 4.1 for more details on each compound.

Table 4.1 Of the 96 compounds classified as hits, 5 were selected for further investigation based on potential promiscuous interactions, homologous ADP1 enzymes, and other literature data

#	m/z	Candidate Name	Preliminary Evidence	Hypothesis for Generation	Relevant ADP1 Genes
1	115.0405	glutarate semialdehyde	• Known reaction exists in *P. putida* (davT) [6] • BLAST hits of P. putida davT, including 4-aminobutyrate aminotransferase • 4-aminobutyrate aminotransferase in ADP1 does a very similar reaction (1 less carbon in chain) [1] • Peak intensity 88x higher in succinate condition [1]; 4-aminobutyrate is 2 steps from succinate [7]	• Promiscuous activity of 4-aminobutyrate aminotransferase on 5-aminopentanoate	• ACIAD3446 - 4-aminobutyrate aminotransferase, PLP-dependent ($E = 10^{-153}$) • ACIAD1210 - diaminobutyrate-2-oxoglutarate transaminase ($E = 10^{-66}$)
2	152.0359	3-hydroxyanthranilate	• Other *Acinetobacter* species can grow on kynurenic acid [2] • BLAST hits of Pseudomonas fluorescens kynurenine 3-monoxygenase (KMO) [4] • Peak seen during quinate (aromatic) feed only [1]	• Unknown kynurenine 3-monoxygenase and/or kynureninase activity	• ACIAD3540 - FAD-dependent urate hydroxylase ($E = 10^{-12}$) • ACIAD0984 - putative hydroxylase involved in salicylate metabolism (SalA-like) ($E = 10^{-7}$) • ACIAD0381 - putative oxidoreductase; putative monooxygenase ($E = 10^{-6}$)
3	161.0460	3-hydroxyadipate	• *E. coli* 3-hydroxyadipyl-CoA dehydrogenase produces a similar compound (with CoA) [5] • BLAST hits of 3-hydroxyadipyl-CoA dehydrogenase • Pimeloyl-CoA synthetase exists in ADP1 and is similar to predicted CoA synthetase reaction [1]	• Promiscuous activity of 3-hydroxylacyl-CoA dehydrogenase creates 3-hydroxyadipyl-CoA • Unknown CoA synthetase activity converts 3-hydroxyadipyl-CoA to 3-hydroxyadipate	• ACIAD1690 - 3-hydroxyacyl-CoA dehydrogenase ($E = 10^{-98}$) • ACIAD2989 - 3-hydroxyacyl-CoA dehydrogenase protein ($E = 10^{-35}$) • ACIAD1826 - pimeloyl-CoA synthetase
4	161.0460	2-ethylmalate	• Predicted reaction previously reported in *S. cerevisiae*, but no enzyme annotated [8] • Very similar reaction (with an extra methyl group) via 2-isopropylmalate synthase in ADP1 [1,9]	• Promiscuous activity of 2-isopropylmalate synthase on 2-oxobutanoate	• ACIAD0530 - 2-isopropylmalate synthase
5	164.0564	4-hydroxy-2,5-dioxopentanoate	• Compound (and predicted precursors) known to bind to YagE, but no reactions reported [10,11] • BLAST hits of E. coli YagE	• Unknown or promiscuous activity of YagE-like enzyme on pyruvate and glyoxal	• ACIAD3585 - dihydrodipicolinate synthase ($E = 10^{-35}$) • ACIAD0130 - 5-dehydro-4-deoxyglucarate dehydratase ($E = 10^{-18}$)

4.3.3.1 Glutarate semialdehyde

The smallest of the five compounds, glutarate semialdehyde (compound 1 in Figure 4.4b), was predicted to be made from either proline, glutaryl-CoA, or glutarate. We found that the predicted reaction from glutarate was in fact a known reaction present in the lysine degradation pathway of *E. coli*, which is not known to exist in ADP1 [140]. Glutarate semialdehyde is a known intermediate of this pathway and is produced from 5-aminopentanoate and 2-oxoglutarate (via glutarate dehydrogenase) and subsequently converted into glutarate (via 5-aminopentanoate aminotransferase) in *E. coli* [140]. We also noticed that the same lysine degradation pathway exists in *Pseudomonas putida*, a soil microbe closely related to ADP1 [28,141,142]. While ADP1 itself contains an annotated glutarate dehydrogenase (ACIAD0131), there is no annotated 5-aminopentanoate aminotransferase in ADP1, so we used BLAST to search for homologs of the *P. putida* 5-aminopentanoate aminotransferase (davT) in the ADP1 genome, resulting in two significant hits. One of these hits coded for 4-aminobutyrate aminotransferase, which is involved in valine degradation and catalyzes the same reaction as 5-aminopentanoate aminotransferase but with one fewer carbon atom in the carbon chain of the substrate containing the amino group. Thus, we hypothesized that the ADP1 4-aminobutyrate aminotransferase may be acting promiscuously on 5-aminopentanoate to produce glutarate semialdehyde. In addition, the unknown peak for this candidate (m/z = 115.0405 Da) had 88 times greater intensity when ADP1 was grown on succinate instead of quinate [38]. This is consistent with this hypothesis, as glutarate semialdehyde is only two steps from succinate in the lysine degradation pathway [140,143].

4.3.3.2 3-hydroxyanthranilate

The second of the five selected compounds, 3-hydroxyanthranilate (compound 2 in Figure 4.4b), is a known biological compound as well, often involved in the kynurenine pathway which degrades kynurenine, a product of tryptophan degradation, and is not known to be in ADP1 [144,145].

We first noticed that one of the predicted reactions was also a reaction reported to occur in this pathway via 3-hydroxyanthranilate dioxygenase which exists in many eukaryotic species including yeast as part of the kynurenine pathway [146]. Some bacteria species are reported to use a kynurenine pathway without a step involving 3-hydroxyanthranilate, suggesting that this enzyme is not widely found in bacteria; the only bacterial species reported to contain this enzyme is *Geobacillus thermodenitrificans* [144,147]. This reaction converts anthranilate into 3-hydroxyanthranilate via anthranilate hydroxylase (GTNG_3160). We used BLAST to search for ADP1 genes homologous to GTNG_3160 but obtained no hits.

However, we also noticed that the known kynurenine pathway in bacteria involves a reaction similar to that of 3-hydroxyanthranilate dioxygenase. Kynurenine-3-monooxygenase (KMO) converts kynurenine to 3-hydroxykynurenine, and so we hypothesized that if ADP1 had a KMO enzyme, it could promiscuously act on anthranilate to produce 3-hydroxyanthranilate. While it is unknown if ADP1 exhibits a kynurenine pathway, other *Acinetobacter* species have been reported to grow on kynurenic acid as a sole carbon source [148]. Additionally, KMO has been identified in a *Pseudomonas* strain, so we used BLAST to search for homologous KMO enzymes in ADP1, obtaining three hits [149]. However, because these hits have relatively high E-values (Table 4.1), they are unlikely to be KMO enzymes, although they may still catalyze KMO-like activity on anthranilate, especially as two of the three hits are not well annotated. Finally, this peak was only observed in the quinate feed condition in the metabolomics dataset. Quinate can be converted by ADP1 to chorismate which is subsequently converted to anthranilate as part of the tryptophan biosynthesis pathway [38,150].

4.3.3.3 3-hydroxyadipate and 2-ethylmalate

The next two compounds were both candidates for the same peak (m/z = 161.0460). They are isomers, differing only by the position of an ethyl group, so they likely have similar retention

times. While normally only one candidate is annotated per peak, due to these similarities, we thought that both warranted further investigation.

The first compound, 3-hydroxyadipate (compound 3 in Figure 4.4b), was predicted via two routes. The first route was a reduction of the carbonyl in beta-ketoadipate (a product of ADP1 aromatic degradation) to a hydroxy group. However, we did not find any dehydrogenases that acted on a similar substrate, so if this is occurring, it would likely be via an unannotated dehydrogenase enzyme. The second predicted route was the hydrolysis of 3-hydroxyadipyl-CoA. This route was at first puzzling, as while 3-hydroxyadipyl-CoA is known to exist in ADP1, only one reaction involving this compound was annotated in BioCyc, producing trans-2,3-dehydroadipyl-CoA, and this product was not annotated as participating in any further reactions. Because we noticed that *E. coli* converts 3-oxoadipyl-CoA (an intermediate in ADP1 aromatic metabolism) into 3-hydroxyadipyl-CoA via 3-hydroxyadipyl-CoA dehydrogenase, we used BLAST to search for homologs in ADP1 [151]. We found two highly significant hits, both annotated as 3-hydroxyacyl-CoA dehydrogenases, which are involved in fatty acid oxidation. If one of these hits converts 3-oxoadipyl-CoA into 3-hydroxyadipyl-CoA in ADP1, then would imply that 3-hydroxyadipyl-CoA is present during growth on aromatics such as quinate. We then noticed that many enzymes perform similar reactions as the predicted second step, where Coenzyme A is replaced with a hydroxy group. We found that pimeloyl-CoA synthetase, which exists in ADP1, had the most similar substrate (pimelate), so we hypothesized that it may be acting promiscuously on a 3-hydroxyadipyl-CoA that is potentially present in ADP1. However, there are likely other unannotated CoA synthetases in ADP1 as well.

The second candidate compound for this peak was 2-ethylmalate (compound 4 in Figure 4.4b). The predicted reaction was previously reported in *S. cerevisiae* but no enzyme was annotated [152]. However, we noticed that a very similar reaction is known to occur in ADP1 via 2-isopropylmalate synthase (LeuA), an enzyme involved in branched chain amino acid biosynthesis

[153]. Because 2-isopropylmalate only differs from 2-ethylmalate by a single methyl group and because there is no known enzyme acting on 2-ethylmalate, we hypothesized that it could be produced via promiscuous activity of the native 2-isopropylmalate synthase in ADP1.

4.3.3.4 4-hydroxy-2,5-dioxopentanoate

The last of the five selected candidates was 4-hydroxy-2,5-dioxopentanoate (compound 5 in Figure 4.4b). While not a known biological compound, we noticed that it existed in the Protein Data Bank (PDB ID: 4OE7) as a ligand for YagE, an *E. coli* gene that is putatively annotated as an aldolase involved in glucose metabolism via the Entner-Doudoroff (ED) pathway[154]. YagE natively ligates glycoaldehyde and pyruvate together to form 2-dehydro-3-deoxyarabinonate[154]. The analogous reaction producing 4-hydroxy-2,5-dioxopentanoate would require glyoxal and pyruvate which were also found to bind to YagE. Because the ED pathway, glyoxal, and pyruvate all exist in ADP1, we searched for YagE homologs via BLAST, hypothesizing that there may be a YagE-like enzyme in ADP1 catalyzing this transformation[20]. This resulted in two significant hits, dihydrodipicolinate synthase and a probable 5-dehydro-4-deoxyglucarate dehydratase. While the annotated reactions for both of these enzymes are not very similar to that of YagE, they may be misannotated (or be highly promiscuous) as a BLAST search of the ADP1 dihydrodipicolinate synthase against the UniProt database resulted in a significant hit ($E = 10^{-172}$) against an enzyme annotated as 4-hydroxy-2-oxoglutarate aldolase in *Lipotes vexillifer* which acts on glyoxylate and pyruvate, very similar to the YagE reaction. Therefore, either (1) these enzymes are misannotated (in ADP1 or in other organisms), (2) YagE-like activity simply does not exist in ADP1, or (3) there is some promiscuous activity of one of these enzymes, which has been reported for related enzymes such as 2-keto-3-deoxygluconate aldolase [155].

4.3.4 Experimental Validation of Predicted Metabolites

Standards of 3-hydroxyanthranilic acid and 3-hydroxyadipic acid were analyzed via LC-QTOF and detected at retention times of 1.73 and 4.14 minutes, respectively. The primary ions detected for each analyte were 152.0356 and 153.0380 for 3-hydroxyanthranilic acid and 161.0457 and 162.0489 for 3-hydroxyadipic acid. Analysis of the ADP1 wild type culture extracts revealed the presence of 3-hydroxyadipic acid at an average concentration of 82.6 μg/L in the extract sample. 3-hydroxyadipic acid was also detected in the ADP1ΔACIAD1826 extracts at an average concentration of 58.12 μg/L. 3-Hydroxyanthranilic acid was not detected in the wild type or knock-out cultures.

4.4 Discussion and Conclusions

This workflow addressed the need for more intelligent candidate set selection and filtering, especially when searching for novel metabolites. By taking an organism- and context-specific approach, candidates were predicted and filtered while maintaining consistency with both the organism's genome-scale metabolic model and the unique features of the metabolomics dataset. While previous work had reported on the prediction of biochemical reactions for candidate set engineering, we (1) considered both the reactome and genome of a specific organism during rule selection, (2) utilized an extended metabolic model (EMM) to help filter out unlikely candidates, and (3) did not use MS2 spectra during candidate ranking [106,112,118].

By starting with a comprehensive set of enzymatic reaction rules, each associated with specific enzymes, we were able to select for both rules that mapped to a reaction in the ADP1 metabolic model as well as rules whose enzymes looked homologous to ADP1 genes. By mapping to both reactions and genes, we obtained a set of known but unknown chemistries existing in ADP1, allowing for the prediction of both promiscuous reactions as well as more chemically novel reactions. This is especially important because one limitation of metabolic

models is that they can be incomplete, especially compared to an organism's genome which tends to be near 100% complete, even for non-model organisms. We validated this gene-based method by demonstrating significant overlap between the reaction-mapped and gene-mapped rule sets. Overall, this combined approach resulted in a more targeted candidate set for ADP1 than using less-organism specific approaches like MINE.

After generating a candidate set using reaction rules specific to ADP1, we further filtered it down by utilizing non-metabolomics information about each candidate. This was especially important in this case as MS2 spectra were not available for ranking candidates for each peak. Metabolic modeling-based features turned out to be highly important for candidate classification. The metabolic model used was the ADP1 EMM that we built, containing all predicted reactions, including gapfilling reactions. Specifically, we found that the distance between a candidate and feed compound in the metabolic model as well as the number of known compounds predicted to carry flux were all good predictors for peak annotation. Because metabolic networks are vast, it is not surprising that candidates closer to feed compounds in metabolic network space were found to be more likely. The number of known compounds predicted to carry flux being important is also consistent with evidence that pathways known to be active based on the metabolomics dataset are most likely to produce other compounds that were measured [156]. We also found that the number of unique precursors for a candidate was an important feature, highlighting the idea that candidates that are predicted to be highly interconnected with known metabolism (which is highly interconnected itself) are more likely [157,158]. Integration of EMM-based metabolic modeling with network-based metabolite identification workflows such as GNPS and *mummichog* could be an area ripe for future development [156,159]. Predicted reaction feasibility via deepRFC was also found to be important for classification [128]. We expect that feasibility prediction tools such as deepRFC will only continue to improve in the future, improving their utility in these types of applications. Finally, we found that our retention time predictor was useful as well, even though the predictions

were not high-resolution, demonstrating that even low quality retention time predictions can still be used to filter out unlikely candidates.

Another important contribution of this work was the fact that MS2 spectra were not required. While highly useful for ranking compounds, as shown by others, MS2 is often not measured for many peaks or even entire datasets, especially when ultra-high-resolution MS1 techniques are used [113–116]. Thus, this workflow can be used as an initial filter to design more targeted metabolomics experiments which could include MS2 for compound identification. Additionally, if a dataset does contain MS2 spectra, the similarity between the measured spectra for a peak and the predicted spectra for a candidate could simply be added as a feature in this classification workflow, with the learner deciding based on the data how much emphasis to put on MS2 spectra versus other non-metabolomics features.

One of the most significant results of this work is that the classifier correctly filtered out 96% of negative candidates for known peaks while retaining 48% of known compounds. Applying this classifier to unknown peaks reduced that candidate set by 98%, greatly reducing the number of compounds to a much more manageable size for further investigation and potential experimental validation. In addition, the classifier could be tuned to further reduce the candidate set size, sacrificing recall to obtain better precision, by increasing the stringency of the classification threshold for labeling a candidate as a hit.

This work demonstrates that cheminformatics tools combined with metabolic modeling and organismal context can be used to generate more targeted and credible candidate sets for metabolite annotation. Our organism-specific approach resulted in a 34-fold reduction in candidate set size compared to using a fully generalized approach and a 3-fold reduction compared to using organism-specific starting compounds with a generalized ruleset. The validated random forest classifier combined information calculated from metabolic modeling, predicted reaction feasibility, and the metabolomics dataset to further reduce this candidate space

by 98% to only the most likely candidates, illustrating that non-metabolomics data can be useful for metabolite annotation. We also experimentally validated 1 compound, 3-hydroxyadipic acid, providing further evidence that this method effectively supports experimental metabolite identification efforts and can aid in improving our knowledge of metabolism.

5. Recommendations for future work

Though the research described in this thesis represents essential progress towards establishment of ADP1 as a industrial chassis organism for lignin upgrading, critical needs should be addressed in future work. In particular:

1. Enhancing oxygen transfer to maintain high ADP1 growth rates while avoiding oxygen starvation
2. Overcoming dilution effects resulting from lowly solubility lignin-derived substrates with low toxicity thresholds
3. Utilization of more complex and realistic substrate streams to reflect industrially relevant feedstocks
4. Leveraging the finer process control available in bioreactor cultivation to explore nitrogen fixation and iron sequestration in co-culture
5. Metabolic engineering towards enhanced storage capacity of cyanophycin peptide granules

5.1 Enhancing oxygen transfer to maintain high ADP1 growth rates while avoiding oxygen starvation

Oxygen transfer can be improved by mechanical and metabolic engineering means. Mechanical methods offer a relatively low risk strategy to improve oxygen transfer using strategies that are ubiquitous in current industrial production platforms [160–162]. The currently available bioreactors in Tyo lab use a ring sparger, which is easily maintained and inexpensive but produces large bubbles results in reduced oxygen transfer surface area. We partially account for this by impeller placement directly above the sparging ring to immediately disperse bubbles and reduce their size. Increasing agitation rates also further disrupts surface tension of bubbles, resulting in increased oxygen transfer rates. However, future strategies could make use of sparging stones, which extrude smaller bubbles via pores and increase oxygen transfer surface area. Furthermore,

my data indicate that shear forces from the current axial flow impellers have a negligible impact on cell viability at agitation rates of 600 – 1200 rpm. ADP1 cells have been demonstrated to possess thick exopolysaccharide capsules, which could contribute to their physical robustness [163]. Determining allowable agitation rates and shear forces for ADP1 in bioreactor cultivation will be vital to future efforts for scale up and commercialization.

Metabolic engineering can also be leveraged to improve oxygen transfer. Specifically, the expression of hemoglobin can improve oxygen uptake. Heterologous expression of bacterial hemoglobin has proven effective at improving growth and productivity in multiple organisms including the obligate aerobic yeast, *Yarrowia lipolytica*. Expression of hemoglobin from *Vitreoscilla* (VHb), an obligate aerobe native to poorly oxygenated environments, has been explored for its growth enhancing properties across several organisms [164–169].

Though the mechanisms of VHb have yet to be fully elucidated, its expression has been demonstrated to improve oxygen diffusion in host cells, resulting in improved growth and production of native and heterologous proteins. When expressed in *E. coli*, VHb confers improved growth rate and overall biomass yields under oxygen limitation [166] and improves synthesis of poly(β-hydroxybutyrate) [168]. Expression of VHb in *Yarrowia lipolytica* improves growth and synthesis of erythritol [164,165]. A primary challenge identified in this thesis is that of supplying adequate oxygen to support rapid growth of ADP1 while minimizing the implementation of expensive oxygenation strategies. In ADP1, expression of VHb holds the potential to generate a superior industrial chassis strain with reduced oxygenation requirements and improved production traits.

Notably, this adaptation also holds the potential to unlock higher throughput methodology for tracking online growth of ADP1. Characterization efforts for ADP1's aromatic acid tolerance, iron sequestration, and nitrogen assimilation in Tyo lab have required higher throughput methodology than is available in shake flask or culture tube growth. However, due to the poor oxygen transfer

in plate-based cultivations, growth rate-based plate assays have been unreliable due to growth rate being limited primarily by oxygen transfer even at very low densities. Growth rate based assays have therefore primarily been conducted either for only low density cells or using more labor-intensive and lower throughput culture conditions. Engineered ADP1 cells expressing VHb could be leveraged to study growth dynamics for a wide array of conditions with higher throughput plate assays.

5.2 Overcoming dilution effects resulting from lowly solubility lignin-derived substrates with low toxicity thresholds

A major hurdle in the upgrading of lignocellulosic feedstocks is the low solubility of many aromatic acid components found in lignin-derived waste streams [68,170]. Heterogeneous waste streams typically contain lowly soluble aromatic acids, which, when used as a primary carbon source, require either economically infeasible processing steps to concentrate feed streams or strategies to prevent dilution from impacting cell growth. High dilution rates during cultivation results in significant loss of biomass in bioreactor effluent, which is resource inefficient and prohibitively expensive. One potential solution to this challenge is the use of cell recyclers, which separate cells from effluent and recycle them to the stirred tank. For ADP1, the use of cell recyclers has not been explored in scaled-up processes but represents a promising solution to this obstacle.

In Tyo lab, several approaches to prevent the continuous production and discard of viable cells may be applied either with current instrumentation or with further investment. Rudimentary strategies for cell recycle involve intermittently separating cells from effluent manually through the use of a centrifuge or some similar cell-sedimentation or filtration device. Previously, this technique has been used to increase volumetric productivity of *Gluconobacter oxydans* in the

synthesis of xylonic acid [171]. Though this methodology represents an accessible strategy for cell recycle, it is labor intensive and infeasible at larger scale without automation.

Alternatively, in-line filtration improves product yields as well as eliminating the need for manual manipulation of bioreactor effluent. An automated and integrated cell recycle loop enables continuous cell recycling with removal of spent medium and is a more realistic strategy for large scale process engineering. Hua *et al.* applied ultrafiltration in tandem with a peristaltic pump to recycle cells back into CSTR while removing filtrate [172]. Cross filtration using hollow fibers has also proven an effective method for cell recycle in *Yarrowia lipolytica* towards improving production of internal lipid products [173], which is particularly of interest in the case of ADP1 due to its highly efficient and native production of wax esters, which form inclusion bodies in cells [32,33,53]. Cell recycle strategies would also enable implementation of continuous phase cultivation, which is a complimentary approach to FBA, a common hypothesis generation tool for metabolic engineering. This also unlocks the ability to more quickly iterate through Design-Build-Test-Cycles, a critical need for leveraging ADP1's unique suitability as an industrial chassis strain. Lastly, removal of filtrate would reduce osmotic stress resulting from continual addition of salt-containing nutrient and carbon feeds, potentially improving long term cell viability and productivity.

5.3 Utilization of more complex and realistic substrate streams to reflect industrially relevant feedstocks

A tangential approach to reducing dilution can leverage ADP1's native catabolic potential for utilizing diverse feedstocks to convert heterogeneous waste streams with higher carbon concentrations to biomass and products. Currently, research has explored ADP1's capability to upgrade a broad range of lignin-derived aromatic acids, glucose, and acetate among other organic molecules [5,23,117,170].

Future efforts could focus on implementing more realistic feedstocks for production strains that have been previously developed in Tyo lab (mevalonate, vanillin glucoside, methyl anthranilate) or could utilize ADP1's diverse metabolism to implement feeding strategies tuned to specific products.

Carbon sources that are metabolically close to the TCA cycle could be converted to acetyl-CoA derived products like wax esters with a relatively low enzymatic burden. Pure acetic acid is one such feedstock and has been utilized as a carbon source for *Y. lipolytica* with optimization of feeding based on measurable process parameters (oxygen uptake rate, carbon dioxide production, and pH) that could be easily applied to ADP1 [173]. Sodium acetate is also efficiently utilized by ADP1 as either a sole carbon source or as a major component of mixed carbon sources and is significantly more soluble than many aromatic acids [20,33,34,36].

Previous engineering has also improved growth of ADP1 on glycolytic substrates and could be applied to enable more efficient upgrading of mixed sugar and aromatic feedstocks [174]. ADP1 glucose catabolism proceeds via the pentose phosphate pathway or the Entner-Doudoroff pathway, which generates pyruvate and glyceraldehyde-3-phosphate (G3P). In some microbes such as *E. coli*, G3P is converted to phosphoenolpyruvate (PEP) and then pyruvate by the enzyme pyruvate kinase. ADP1 lacks pyruvate kinase, which increases the enzymatic burden of transforming G3P to pyruvate [175]. Natively, this results in significant flux of glycolytic substrates towards biosynthetic functions, such as production of exopolysaccharide synthase [163,176]. Kannisto *et al.* expressed *pykF* (*E. coli)* in ADP1 to enable direct conversion of PEP to pyruvate. This simple engineering step improved growth on glucose and gluconate, both of which can be present in lignocellulose-derived feedstocks, at times even as the primary component [68,170,174]. Efficient utilization of inexpensive, mixed aromatic and sugar feedstocks can significantly reduce process costs, which is a vital step in industrialization of ADP1.

5.4 Leveraging the finer process control available in bioreactor cultivation to explore nitrogen fixation and iron sequestration in co-culture

The development of growth strategies for ADP1-based nitrogen sequestration in non-sterile environments would greatly benefit from exploring consortium behavior in bioreactors. A current emphasis area in Tyo lab is ADP1's nitrogen sequestration capabilities for the denitrification of non-sterile waste streams. By converting nitrogen to the storage polymer, cyanophycin (CPG), ADP1 efficiently denitrifies waste streams, which is a major challenge in wastewater treatment [177–180]. Iron sequestration in ADP1 consortia is also a current focus in Tyo lab, with a particular focus on artificial consortia dynamics using ADP1 and *E. coli*. Initial work, both with nitrogen fixation and iron sequestration, has primarily been conducted in shake flask and plate-based growth environments, which is ideal for medium- to high-throughput screening. However, these cultivation settings are limited by both control of process variables and by informational richness based on available online signals.

The available online measurements and control of key process variables—especially oxygen transfer and pH—in bioreactor cultivation will be essential to achieve long-term research goals of understanding consortia dynamics and developing methods to leverage ADP1's nutrient sequestration abilities. Chemostat cultivation, which is typical of wastewater treatment strategies [181–183], can be implemented at lab scale using rudimentary equipment. For example, the sampling port of our current bioreactor system can be run through a peristaltic pump, thus enabling continuous removal of culture broth at the dilution rate. Implementing continuous cultivation also enables steady state metabolic flux measurements for use in flux balance analysis, which in turn can elucidate metabolic engineering steps for strain improvement towards specific goals such as enhancing ADP1's capability for CPG-based nitrogen fixation.

5.5 Metabolic engineering towards enhanced storage capacity of cyanophycin peptide granules (CPG)

Lastly, a recent development in enhancing ADP1's wax ester production—the use of morphologically engineered cells to increase ADP1's cell volume—may also hold promise for production of CPG, another intracellular product. Due to the nature of intracellular product synthesis and storage, production capacity is limited by cell morphology. Recently, Luo *et al.* leveraged CRISPR interference to reshape ADP1 cells, increasing their volume and thus wax ester and other internal product storage capacity [79]. Increased cell size reduces the carbon cost of denitrification by improving the CPG production yield per cell. It also reduces downstream processing costs by reducing the cost of cell separation [184]. Thus, the application of this strategy to CPG synthesis has the potential to enhance production of CPG and reduce overall process costs.

6. References

1. Nicholson, S. R., Rorrer, N. A., Carpenter, A. C. & Beckham, G. T. Manufacturing energy and greenhouse gas emissions associated with plastics consumption. *Joule* **5**, 673–686 (2021).
2. Fura, A. *et al.* Discovering drugs through biological transformation: Role of pharmacologically active metabolites in drug discovery. *J. Med. Chem.* **47**, 4339–4351 (2004).
3. Bradshaw, M. J. The geopolitics of global energy security. *Geogr. Compass* **3**, 1920–1937 (2009).
4. Xu, Z., Lei, P., Zhai, R., Wen, Z. & Jin, M. Recent advances in lignin valorization with bacterial cultures: Microorganisms, metabolic pathways, and bio-products. *Biotechnol. Biofuels* **12**, 1–19 (2019).
5. Borchert, A. J., Henson, W. R. & Beckham, G. T. Challenges and opportunities in biological funneling of heterogeneous and toxic substrates beyond lignin. *Curr. Opin. Biotechnol.* **73**, 1–13 (2022).
6. José Borges Gomes, F., de Souza, R. E., Brito, E. O. & Costa Lelis, R. C. A review on lignin sources and uses. *J. Appl. Biotechnol. Bioeng.* **7**, 100–105 (2020).
7. Wang, H., Pu, Y., Ragauskas, A. & Yang, B. From lignin to valuable products–strategies, challenges, and prospects. *Bioresour. Technol.* **271**, 449–461 (2019).
8. Perlack, RD; Stokes, B. US Billion Ton Update: Biomass supply for a bioenergy and bioproducts industry (executive summary). *Oak Ridge US Dep. Energy, Oak Ridg* **7**, 375–380 (2011).
9. Chen, Z. & Wan, C. Biological valorization strategies for converting lignin into fuels and chemicals. *Renew. Sustain. Energy Rev.* **73**, 610–621 (2017).
10. Ragauskas, A. J. *et al.* Lignin valorization: Improving lignin processing in the biorefinery. *Science (80-.).* **344**, (2014).
11. Zakzeski, J., Bruijnincx, P. C. A., Jongerius, A. L. & Weckhuysen, B. M. The Catalytic Valorization of Lignin for the Production of Renewable Chemicals. *Chem. Rev.* **110**, 3552–3599 (2010).
12. 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 Chem.* **17**, 4951–4967 (2015).
13. Bugg, T. D. H., Ahmad, M., Hardiman, E. M. & Rahmanpour, R. Pathways for degradation of lignin in bacteria and fungi. *Nat. Prod. Rep.* **28**, 1883–1896 (2011).
14. Pollegioni, L., Tonin, F. & Rosini, E. Lignin-degrading enzymes. *FEBS J.* **282**, 1190–1213 (2015).
15. Kamimura, N., Sakamoto, S., Mitsuda, N., Masai, E. & Kajita, S. Advances in microbial lignin degradation and its applications. *Curr. Opin. Biotechnol.* **56**, 179–186 (2019).
16. Keasling, J. D. Synthetic biology for synthetic chemistry. *ACS Chem. Biol.* **3**, 64–76 (2008).
17. Davis, K. & Moon, T. S. Tailoring microbes to upgrade lignin. *Curr. Opin. Chem. Biol.* **59**, 23–29 (2020).
18. Wu, W., Liu, F. & Singh, S. Toward engineering E. coli with an autoregulatory system for lignin valorization. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2970–2975 (2018).
19. Wu, W. *et al.* Lignin Valorization: Two Hybrid Biochemical Routes for the Conversion of Polymeric Lignin into Value-added Chemicals. *Sci. Rep.* **7**, 1–13 (2017).
20. Kannisto, M. S. *et al.* Metabolic engineering of Acinetobacter baylyi ADP1 for removal of Clostridium butyricum growth inhibitors produced from lignocellulosic hydrolysates. *Biotechnol. Biofuels* **8**, 198 (2015).

21. Singh, A. *et al.* Removal of aromatic inhibitors produced from lignocellulosic hydrolysates by Acinetobacter baylyi ADP1 with formation of ethanol by Kluyveromyces marxianus. *Biotechnol. Biofuels* **12**, 1–11 (2019).
22. Clarkson, S. M. *et al.* Construction and Optimization of a Heterologous Pathway for Protocatechuate Catabolism in Escherichia coli Enables Bioconversion of Model Aromatic Compounds. *Appl. Environ. Microbiol.* **83**, e01313-17 (2017).
23. Young, D. M., Parke, D. & Ornston, L. N. OPPORTUNITIES FOR GENETIC INVESTIGATION AFFORDED BY ACINETOBACTER BAYLYI, A NUTRITIONALLY VERSATILE BACTERIAL SPECIES THAT IS HIGHLY COMPETENT FOR NATURAL TRANSFORMATION. *Annu. Rev. Microbiol.* **59**, 519–551 (2005).
24. Metzgar, D. *et al.* Acinetobacter sp. ADP1: An ideal model organism for genetic analysis and genome engineering. *Nucleic Acids Res.* **32**, 5780–5790 (2004).
25. Elliott, K. T. & Neidle, E. L. Acinetobacter baylyi ADP1: Transforming the choice of model organism. *IUBMB Life* **63**, 1075–1080 (2011).
26. Wells, T. & Ragauskas, A. J. Biotechnological opportunities with the β-ketoadipate pathway. *Trends Biotechnol.* **30**, 627–637 (2012).
27. Vaneechoutte, M. *et al.* Naturally transformable Acinetobacter sp. strain ADP1 belongs to the newly described species Acinetobacter baylyi. *Appl. Environ. Microbiol.* **72**, 932–936 (2006).
28. Barbe, V. *et al.* Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Res.* **32**, 5766–5779 (2004).
29. Biggs, B. W. *et al.* Development of a genetic toolset for the highly engineerable and metabolically versatile Acinetobacter baylyi ADP1. *Nucleic Acids Res.* **48**, 5169–5182 (2020).
30. Juni, E. & Janik, A. Transformation of Acinetobacter calco-aceticus (Bacterium anitratum). *J. Bacteriol.* **98**, 281–288 (1969).
31. Suárez, G. A. *et al.* Rapid and assured genetic engineering methods applied to Acinetobacter baylyi ADP1 genome streamlining. *Nucleic Acids Res.* **48**, 4585–4600 (2020).
32. Santala, S., Efimova, E., Koskinen, P., Karp, M. T. & Santala, V. Rewiring the wax ester production pathway of acinetobacter baylyi ADP1. *ACS Synth. Biol.* **3**, 145–151 (2014).
33. Santala, S., Santala, V., Liu, N. & Stephanopoulos, G. Partitioning metabolism between growth and product synthesis for coordinated production of wax esters in Acinetobacter baylyi ADP1. *Biotechnol. Bioeng.* **118**, 2283–2292 (2021).
34. Kannisto, M., Efimova, E., Karp, M. & Santala, V. Growth and wax ester production of an Acinetobacter baylyi ADP1 mutant deficient in exopolysaccharide capsule synthesis. *J. Ind. Microbiol. Biotechnol.* **44**, 99–105 (2017).
35. Luo, J., Efimova, E., Losoi, P., Santala, V. & Santala, S. Wax ester production in nitrogen-rich conditions by metabolically engineered Acinetobacter baylyi ADP1. *Metab. Eng. Commun.* **10**, (2020).
36. Salmela, M., Lehtinen, T., Efimova, E., Santala, S. & Santala, V. Alkane and wax ester production from lignin-related aromatic compounds. *Biotechnol. Bioeng.* **116**, 1934–1945 (2019).
37. Lehtinen, T., Efimova, E., Santala, S. & Santala, V. Improved fatty aldehyde and wax ester production by overexpression of fatty acyl-CoA reductases. *Microb. Cell Fact.* **17**, 19 (2018).
38. Stuani, L. *et al.* Novel metabolic features in Acinetobacter baylyi ADP1 revealed by a multiomics approach. *Metabolomics* **10**, 1223–1238 (2014).

39. Arvay, E., Biggs, B. W., Guerrero, L., Jiang, V. & Tyo, K. Engineering Acinetobacter baylyi ADP1 for mevalonate production from lignin-derived aromatic compounds. *Metab. Eng. Commun.* **13**, e00173 (2021).
40. Norgren, M. & Edlund, H. Lignin: Recent advances and emerging applications. *Curr. Opin. Colloid Interface Sci.* **19**, 409–416 (2014).
41. Linger, J. G. *et al.* Lignin valorization through integrated biological funneling and chemical catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12013–12018 (2014).
42. Schutyser, W. *et al.* Chemicals from lignin: An interplay of lignocellulose fractionation, depolymerisation, and upgrading. *Chem. Soc. Rev.* **47**, 852–908 (2018).
43. MacEachran, D. P. & Sinskey, A. J. The Rhodococcus opacus TadD protein mediates triacylglycerol metabolism by regulating intracellular NAD(P)H pools. *Microb. Cell Fact.* **12**, 1–12 (2013).
44. Beckham, G. T., Johnson, C. W., Karp, E. M., Salvachúa, D. & Vardon, D. R. Opportunities and challenges in biological lignin valorization. *Curr. Opin. Biotechnol.* **42**, 40–53 (2016).
45. Elliott, K. T. & Neidle, E. L. Acinetobacter baylyi ADP1: Transforming the choice of model organism. *IUBMB Life* **63**, 1075–1080 (2011).
46. Biggs, B. W. *et al.* Development of a genetic toolset for the highly engineerable and metabolically versatile Acinetobacter baylyi ADP1. *Nucleic Acids Res.* 1–14 (2020) doi:10.1093/nar/gkaa167.
47. Suárez, G. A. *et al.* Rapid and assured genetic engineering methods applied to Acinetobacter baylyi ADP1 genome streamlining. (2019) doi:10.1101/754242.
48. Wang, J. Engineering of a highly efficient Escherichia coli strain for mevalonate fermentation through chromosomal integration. *Appl. Environ. Microbiol.* **2**, 1069–1079 (2017).
49. Ajikumar, P. K. *et al.* Terpenoids: Opportunities for biosynthesis of natural product drugs using engineered microorganisms. *Mol. Pharm.* **5**, 167–190 (2008).
50. Belcher, M. S., Mahinthakumar, J. & Keasling, J. D. New frontiers: harnessing pivotal advances in microbial engineering for the biosynthesis of plant-derived terpenoids. *Curr. Opin. Biotechnol.* **65**, 88–93 (2020).
51. Zhang, Y., Nielsen, J. & Liu, Z. Engineering yeast metabolism for production of terpenoids for use as perfume ingredients, pharmaceuticals and biofuels. *FEMS Yeast Res.* **17**, 1–11 (2017).
52. Ishige, T. *et al.* Wax Ester Production from n-Alkanes by Acinetobacter sp. Strain M-1: Ultrastructure of Cellular Inclusions and Role of Acyl Coenzyme A Reductase. *Appl. Environ. Microbiol.* **68**, 1192–1195 (2002).
53. Santala, S. *et al.* Improved triacylglycerol production in Acinetobacter baylyi ADP1 by metabolic engineering. *Microb. Cell Fact.* **10**, 36 (2011).
54. Luo, J., Efimova, E., Losoi, P., Santala, V. & Santala, S. Wax ester production in nitrogen-rich conditions by metabolically engineered Acinetobacter baylyi ADP1. *bioRxiv* 735274 (2019) doi:10.1101/735274.
55. Santala, S., Efimova, E., Koskinen, P., Karp, M. T. & Santala, V. Rewiring the wax ester production pathway of acinetobacter baylyi ADP1. *ACS Synth. Biol.* **3**, 145–151 (2014).
56. Vaneechoutte, M. *et al.* Naturally Transformable Acinetobacter sp. Strain ADP1 Belongs to the Newly Described Species Acinetobacter baylyi. *Appl. Environ. Microbiol.* **72**, 932–936 (2006).
57. Alonso-Gutierrez, J. *et al*. Metabolic engineering of Escherichia coli for limonene and perillyl alcohol production. *Metab. Eng.* **19**, 33–41 (2013).
58. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred

kilobases. *Nat. Methods* **6**, 343–345 (2009).
59. Bryksin, A. V. & Matsumura, I. Rational design of a plasmid origin that replicates efficiently in both gram-positive and gram-negative bacteria. *PLoS One* **5**, (2010).
60. Metzgar, D. *et al.* Acinetobacter sp. ADP1: An ideal model organism for genetic analysis and genome engineering. *Nucleic Acids Res.* **32**, 5780–5790 (2004).
61. Anton V. Bryksin and Ichiro Matsumura. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. **48**, 463–465 (2010).
62. Blomfield, I. C., Vaughn, V., Rest, R. F. & Eisenstein, B. I. Allelic exchange in Escherichia coli using the Bacillus subtilis sacB gene and a temperature-sensitive pSC101 replicon. *Mol. Microbiol.* **5**, 1447–1457 (1991).
63. Dueber, J. E. *et al.* Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* **27**, 753–759 (2009).
64. Garner, B. L., Arceneaux, J. E. L. & Byers, B. R. Temperature control of a 3,4-dihydroxybenzoate (protocatechuate)-based siderophore in Bacillus anthracis. *Curr. Microbiol.* **49**, 89–94 (2004).
65. Fischer, R., Bleichrodt, F. S. & Gerischer, U. C. Aromatic degradative pathways in Acinetobacter baylyi underlie carbon catabolite repression. *Microbiology* **154**, 3095–3103 (2008).
66. Alvarez, H. M. & Steinbüchel, A. Triacylglycerols in prokaryotic microorganisms. *Appl. Microbiol. Biotechnol.* **60**, 367–376 (2002).
67. Salmela, M., Lehtinen, T., Efimova, E., Santala, S. & Santala, V. Alkane and wax ester production from lignin-related aromatic compounds. *Biotechnol. Bioeng.* **116**, 1934–1945 (2019).
68. Beckham, G. T., Johnson, C. W., Karp, E. M., Salvachúa, D. & Vardon, D. R. Opportunities and challenges in biological lignin valorization. *Curr. Opin. Biotechnol.* **42**, 40–53 (2016).
69. Elliott, K. T. & Neidle, E. L. Acinetobacter baylyi ADP1: Transforming the choice of model organism. *IUBMB Life* **63**, 1075–1080 (2011).
70. Xu, Z., Lei, P., Zhai, R., Wen, Z. & Jin, M. Recent advances in lignin valorization with bacterial cultures: microorganisms, metabolic pathways, and bio-products. *Biotechnol. Biofuels* **12**, 32 (2019).
71. Young, D. M., Parke, D. & Ornston, L. N. Opportunities for Genetic Investigation Afforded by Acinetobacter baylyi , A Nutritionally Versatile Bacterial Species that is Highly Competent for Natural Transformation. *Annu. Rev. Microbiol.* **59**, 19–51 (2005).
72. Borchert, A. J., Henson, W. R. & Beckham, G. T. Challenges and opportunities in biological funneling of heterogeneous and toxic substrates beyond lignin. *Curr. Opin. Biotechnol.* **73**, 1–13 (2022).
73. Kannisto, M. S. *et al.* Metabolic engineering of Acinetobacter baylyi ADP1 for removal of Clostridium butyricum growth inhibitors produced from lignocellulosic hydrolysates. *Biotechnol. Biofuels* **8**, 198 (2015).
74. Luo, J. *et al.* Characterization of Highly Ferulate-Tolerant Acinetobacter baylyi ADP1 Isolates by a Rapid Reverse Engineering Method. *Appl. Environ. Microbiol.* **88**, 1–20 (2022).
75. Biggs, B. W. *et al.* Development of a genetic toolset for the highly engineerable and metabolically versatile Acinetobacter baylyi ADP1. *Nucleic Acids Res.* **48**, 5169–5182 (2020).
76. Suárez, G. A. *et al.* Rapid and assured genetic engineering methods applied to acinetobacter baylyi ADP1 genome streamlining. *Nucleic Acids Res.* **48**, 4585–4600 (2021).

77. Luo, J., Efimova, E., Losoi, P., Santala, V. & Santala, S. Wax ester production in nitrogen-rich conditions by metabolically engineered Acinetobacter baylyi ADP1. *Metab. Eng. Commun.* **10**, (2020).
78. Kannisto, M., Efimova, E., Karp, M. & Santala, V. Growth and wax ester production of an Acinetobacter baylyi ADP1 mutant deficient in exopolysaccharide capsule synthesis. *J. Ind. Microbiol. Biotechnol.* **44**, 99–105 (2017).
79. Luo, J., Efimova, E., Volke, D. C., Santala, V. & Santala, S. Engineering cell morphology by CRISPR interference in Acinetobacter baylyi ADP1. *Microb. Biotechnol.* **15**, 2800–2818 (2022).
80. Luo, J., Lehtinen, T., Efimova, E., Santala, V. & Santala, S. Synthetic metabolic pathway for the production of 1-alkenes from lignin-derived molecules. *Microb. Cell Fact.* **18**, 1–13 (2019).
81. Ren, Y., Ling, C., Hajnal, I., Wu, Q. & Chen, G. Q. Construction of Halomonas bluephagenesis capable of high cell density growth for efficient PHA production. *Appl. Microbiol. Biotechnol.* **102**, 4499–4510 (2018).
82. Ankenbauer, A. *et al.* Pseudomonas putida KT2440 is naturally endowed to withstand industrial-scale stress conditions. *Microb. Biotechnol.* **13**, 1145–1161 (2020).
83. Neubauer, P. & Junne, S. Scale-down simulators for metabolic analysis of large-scale bioprocesses. *Curr. Opin. Biotechnol.* **21**, 114–121 (2010).
84. Käß, F. *et al.* Rapid assessment of oxygen transfer impact for Corynebacterium glutamicum. *Bioprocess Biosyst. Eng.* **37**, 2567–2577 (2014).
85. Bauer, S. & Ziv, E. Dense growth of aerobic bacteria in a bench-scale fermentor. *Biotechnol. Bioeng.* **18**, 81–94 (1976).
86. González, R. H., Nusblat, A. & Nudel, B. C. Detection and characterization of quorum sensing signal molecules in Acinetobacter strains. *Microbiol. Res.* **155**, 271–277 (2001).
87. Subhadra, B., Oh, M. H. & Choi, C. H. Quorum sensing in Acinetobacter: With special emphasis on antibiotic resistance, biofilm formation and quorum quenching. *AIMS Microbiol.* **2**, 27–41 (2016).
88. Subhadra, B. *et al.* The osmotic stress response operon betIBA is under the functional regulation of BetI and the quorum-sensing regulator AnoR in Acinetobacter nosocomialis. *J. Microbiol.* **58**, 519–529 (2020).
89. Alvarez, H. & Steinbüchel, A. Triacylglycerols in prokaryotic microorganisms. *Appl. Microbiol. Biotechnol.* **60**, 367–376 (2002).
90. Linger, J. G. *et al.* Lignin valorization through integrated biological funneling and chemical catalysis. *Proc. Natl. Acad. Sci.* **111**, 12013–12018 (2014).
91. Guzmán, G. I. *et al.* Model-driven discovery of underground metabolic functions in Escherichia coli. *Proc. Natl. Acad. Sci.* **112**, 929–934 (2015).
92. Wang, Y. *et al.* Untargeted Metabolomics Identifies Novel Potential Biomarkers of Habitual Food Intake in a Cross-Sectional Study of Postmenopausal Women. *J. Nutr.* **148**, 932–943 (2018).
93. Caldwell, G. W. & Leo, G. C. Can Untargeted Metabolomics Be Utilized in Drug Discovery/Development? *Curr. Top. Med. Chem.* **17**, 2716–2739 (2017).
94. Vavricka, C. J., Hasunuma, T. & Kondo, A. Dynamic Metabolomics for Engineering Biology: Accelerating Learning Cycles for Bioproduction. *Trends Biotechnol.* **38**, 68–82 (2020).
95. Chaleckis, R., Meister, I., Zhang, P. & Wheelock, C. E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr. Opin. Biotechnol.* **55**, 44–50 (2019).
96. Huang, B.-M. *et al.* Discovery of markers for discriminating the age of cultivated ginseng

by using UHPLC-QTOF/MS coupled with OPLS-DA. *Phytomedicine* **45**, 8–17 (2018).
97. Reymond, J.-L. & Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.* **3**, 649–657 (2012).
98. Low, D. Y. *et al.* Data sharing in PredRet for accurate prediction of retention time: Application to plant food bioactive compounds. *Food Chem.* **357**, 129757 (2021).
99. Stanstrup, J., Neumann, S. & Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal. Chem.* **87**, 9421–9428 (2015).
100. Sawada, Y. *et al.* Widely Targeted Metabolomics Based on Large-Scale MS/MS Data for Elucidating Metabolite Accumulation Patterns in Plants. *Plant Cell Physiol.* **50**, 37–47 (2009).
101. Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D. & Mattivi, F. The influence of storage on the “chemical age” of red wines. *Metabolomics* **10**, 816–832 (2014).
102. Cao, M. *et al.* Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **11**, 696–706 (2015).
103. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
104. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).
105. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
106. Jeffryes, J. G. *et al.* MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminform.* **7**, 44 (2015).
107. Sorokina, M. & Steinbeck, C. NaPLeS: a natural products likeness scorer—web application and database. *J. Cheminform.* **11**, 55 (2019).
108. Strutz, J., Shebek, K. M., Broadbelt, L. J. & Tyo, K. E. J. MINE 2.0: enhanced biochemical coverage for peak identification in untargeted metabolomics. *Bioinformatics* **38**, 3484–3487 (2022).
109. Djoumbou-Feunang, Y. *et al.* BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminform.* **11**, 2 (2019).
110. Ni, Z., Stine, A. E., Tyo, K. E. J. & Broadbelt, L. J. Curating a comprehensive set of enzymatic reaction rules for efficient novel biosynthetic pathway design. *Metab. Eng.* **65**, 79–87 (2021).
111. Duigou, T., du Lac, M., Carbonell, P. & Faulon, J.-L. RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Res.* **47**, D1229–D1235 (2019).
112. Menikarachchi, L. C., Hill, D. W., Hamdalla, M. A., Mandoiu, I. I. & Grant, D. F. In Silico Enzymatic Synthesis of a 400 000 Compound Biochemical Database for Nontargeted Metabolomics. *J. Chem. Inf. Model.* **53**, 2483–2492 (2013).
113. Wang, F. *et al.* CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **93**, 11692–11700 (2021).
114. Sévin, D. C., Fuhrer, T., Zamboni, N. & Sauer, U. Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli. *Nat. Methods* **14**, 187–194 (2017).
115. Kao, C. Y., Kuo, P. Y. & Liao, H. W. Untargeted microbial exometabolomics and metabolomics analysis of helicobacter pylori j99 and jhp0106 mutant. *Metabolites* **11**, (2021).
116. Park, K. H. *et al.* Simultaneous molecular formula determinations of natural compounds in

a plant extract using 15 T Fourier transform ion cyclotron resonance mass spectrometry. *Plant Methods* **9**, 15 (2013).
117. Arvay, E., Biggs, B. W., Guerrero, L., Jiang, V. & Tyo, K. Engineering Acinetobacter baylyi ADP1 for mevalonate production from lignin-derived aromatic compounds. *Metab. Eng. Commun.* **13**, e00173 (2021).
118. Hassanpour, N. *et al.* Biological filtering and substrate promiscuity prediction for annotating untargeted metabolomics. *Metabolites* **10**, (2020).
119. Durot, M. *et al.* Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst. Biol.* **2**, 85 (2008).
120. Li, C. *et al.* BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* **4**, 92 (2010).
121. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **84**, 5035–5039 (2012).
122. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
123. Landrum, G. RDKit, Open-Source Cheminformatics.
124. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
125. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
126. Jeffryes, J. G. *et al.* Chemical-damage MINE: A database of curated and predicted spontaneous metabolic reactions. *Metab. Eng.* **69**, 302–312 (2022).
127. Amin, S. A., Chavez, E., Porokhin, V., Nair, N. U. & Hassoun, S. Towards creating an extended metabolic model (EMM) for E. coli using enzyme promiscuity prediction and metabolomics data. *Microb. Cell Fact.* **18**, 109 (2019).
128. Kim, Y., Ryu, J. Y., Kim, H. U., Jang, W. D. & Lee, S. Y. A deep learning approach to evaluate the feasibility of enzymatic reactions generated by retrobiosynthesis. *Biotechnol. J.* **16**, 2000605 (2021).
129. Steyerberg, E. W. Applications of prediction models BT - Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. in (ed. Steyerberg, E. W.) 11–31 (Springer New York, 2009). doi:10.1007/978-0-387-77244-8_2.
130. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
131. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
132. Zhang, X., Gweon, H. & Provost, S. Threshold Moving Approaches for Addressing the Class Imbalance Problem and their Application to Multi-label Classification. *ACM Int. Conf. Proceeding Ser.* **PartF16925**, 72–77 (2020).
133. Brunzell, H. & Eriksson, J. Feature reduction for classification of multidimensional data. *Pattern Recognit.* **33**, 1741–1748 (2000).
134. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* **49**, D498–D508 (2021).
135. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).
136. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
137. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
138. Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application

for prioritization of compound libraries. *J. Chem. Inf. Model.* **48**, 68–74 (2008).
139. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).
140. Knorr, S. *et al.* Widespread bacterial lysine degradation proceeding via glutarate and L-2-hydroxyglutarate. *Nat. Commun.* **9**, 5071 (2018).
141. Zhang, M. *et al.* Increased glutarate production by blocking the glutaryl-CoA dehydrogenation pathway and a catabolic pathway involving l-2-hydroxyglutarate. *Nat. Commun.* **9**, 2114 (2018).
142. Olga, R., Manuel, E.-U., Soeren, M. & L., R. J. The davDT Operon of Pseudomonas putida, Involved in Lysine Catabolism, Is Induced in Response to the Pathway Intermediate δ-Aminovaleric Acid. *J. Bacteriol.* **186**, 3439–3446 (2004).
143. Sarasa, S. B. *et al.* A Brief Review on the Non-protein Amino Acid, Gamma-amino Butyric Acid (GABA): Its Production and Role in Microbes. *Curr. Microbiol.* **77**, 534–544 (2020).
144. Kurnasov, O. *et al.* Aerobic tryptophan degradation pathway in bacteria: novel kynurenine formamidase. *FEMS Microbiol. Lett.* **227**, 219–227 (2003).
145. Bortolotti, P. *et al.* Tryptophan catabolism in Pseudomonas aeruginosa and potential for inter-kingdom relationship. *BMC Microbiol.* **16**, 137 (2016).
146. Kucharczyk, R., Zagulski, M., Rytka, J. & Herbert, C. J. The yeast gene YJR025c encodes a 3-hydroxyanthranilic acid dioxygenase and is involved in nicotinic acid biosynthesis. *FEBS Lett.* **424**, 127–130 (1998).
147. Liu, X. *et al.* Characterization of the anthranilate degradation pathway in Geobacillus thermodenitrificans NG80-2. *Microbiology* **156**, 589–595 (2010).
148. Breisch, J., Huber, L. S., Kraiczy, P., Hubloher, J. & Averhoff, B. The ß-ketoadipate pathway of Acinetobacter baumannii is involved in complement resistance and affects resistance against aromatic antibiotics. *Environ. Microbiol. Rep.* **14**, 170–178 (2022).
149. Crozier-Reabe, K. R., Phillips, R. S. & Moran, G. R. Kynurenine 3-Monooxygenase from Pseudomonas fluorescens: Substrate-like Inhibitors both Stimulate Flavin Reduction and Stabilize the Flavin–Peroxo Intermediate yet Result in the Production of Hydrogen Peroxide. *Biochemistry* **47**, 12420–12433 (2008).
150. Radwanski, E. R. & Last, R. L. Tryptophan Biosynthesis and Molecular Genetics Biochemical - Regulation of the Tryptophan Biosynthetic Pathway in Plants. *Plant Cell* **7**, 929 (1995).
151. Teufel, R. *et al.* Bacterial phenylalanine and phenylacetate catabolic pathway revealed. *Proc. Natl. Acad. Sci.* **107**, 14390–14395 (2010).
152. Strassman, M. & Ceci, L. N. A study of acetyl-CoA condensation with α-keto acids. *Arch. Biochem. Biophys.* **119**, 420–428 (1967).
153. Frantom, P. A. Structural and functional characterization of α-isopropylmalate synthase and citramalate synthase, members of the LeuA dimer superfamily. *Arch. Biochem. Biophys.* **519**, 202–209 (2012).
154. Bhaskar, V. *et al.* Identification of biochemical and putative biological role of a xenolog from Escherichia coli using structural analysis. *Proteins Struct. Funct. Bioinforma.* **79**, 1132–1142 (2011).
155. Theodossis, A. *et al.* The structural basis for substrate promiscuity in 2-keto-3-deoxygluconate aldolase from the Entner-Doudoroff pathway in Sulfolobus solfataricus. *J. Biol. Chem.* **279**, 43886–43892 (2004).
156. Li, S. *et al.* Predicting Network Activity from High Throughput Metabolomics. *PLOS Comput. Biol.* **9**, e1003123 (2013).
157. Pfeiffer, T., Soyer, O. S. & Bonhoeffer, S. The Evolution of Connectivity in Metabolic Networks. *PLOS Biol.* **3**, e228 (2005).

158. Guimerà, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
159. Aron, A. T. *et al.* Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
160. Amer, M., Feng, Y. & Ramsey, J. D. Using CFD simulations and statistical analysis to correlate oxygen mass transfer coefficient to both geometrical parameters and operating conditions in a stirred-tank bioreactor. *Biotechnol. Prog.* **35**, 1–14 (2019).
161. Garcia-Ochoa, F. & Gomez, E. Bioreactor scale-up and oxygen transfer rate in microbial processes: An overview. *Biotechnol. Adv.* **27**, 153–176 (2009).
162. Karimi, A. *et al.* Oxygen mass transfer in a stirred tank bioreactor using different impeller configurations for environmental purposes. *J. Environ. Heal. Sci. Eng.* **10**, 1–9 (2013).
163. Kaplan, N. & Rosenberg, E. Exopolysaccharide distribution of and bioemulsifier production by Acinetobacter calcoaceticus BD4 and BD413. *Appl. Environ. Microbiol.* **44**, 1335–1341 (1982).
164. Bhave, S. L. & Chattoo, B. B. Expression of Vitreoscilla Hemoglobin Improves Growth and Levels of Extracellular Enzyme in Yarrowia lipolytica. *Biotechnol. Bioeng.* **84**, 658–666 (2003).
165. Mirończuk, A. M. *et al.* Heterologous overexpression of bacterial hemoglobin VHb improves erythritol biosynthesis by yeast Yarrowia lipolytica. *Microb. Cell Fact.* **18**, 1–8 (2019).
166. Khosla, C. & Bailey, J. E. Heterologous expression of a bacterial haemoglobin improves the growth properties of recombinant Escherichia coli. *Nature* **331**, 633–635 (1988).
167. Frey, A. D. & Kallio, P. T. Bacterial hemoglobins and flavohemoglobins: Versatile proteins and their impact on microbiology and biotechnology. *FEMS Microbiol. Rev.* **27**, 525–545 (2003).
168. Yu, H., Shi, Y., Zhang, Y., Yang, S. & Shen, Z. Effect of Vitreoscilla hemoglobin biosynthesis in Escherichia coli on production of poly(β-hydroxybutyrate) and fermentative parameters. *FEMS Microbiol. Lett.* **214**, 223–227 (2002).
169. Zhang, L. *et al.* Recent developments and future prospects of Vitreoscilla hemoglobin application in metabolic engineering. *Biotechnol. Adv.* **25**, 123–136 (2007).
170. Karp, E. M. *et al.* Quantification of acidic compounds in complex biomass-derived streams. *Green Chem.* **18**, 4750–4760 (2016).
171. Zhou, X., Zhou, X. & Xu, Y. Improvement of fermentation performance of Gluconobacter oxydans by combination of enhanced oxygen mass transfer in compressed-oxygen-supplied sealed system and cell-recycle technique. *Bioresour. Technol.* **244**, 1137–1141 (2017).
172. Hua, X., Cao, R., Zhou, X. & Xu, Y. One-step continuous/semi-continuous whole-cell catalysis production of glycolic acid by a combining bioprocess with in-situ cell recycling and electrodialysis. *Bioresour. Technol.* **273**, 515–520 (2019).
173. Xu, J., Liu, N., Qiao, K., Vogg, S. & Stephanopoulos, G. Application of metabolic controls for the maximization of lipid production in semicontinuous fermentation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5308–E5316 (2017).
174. Kannisto, M., Aho, T., Karp, M. & Santala, V. Metabolic engineering of Acinetobacter baylyi ADP1 for improved growth on gluconate and glucose. *Appl. Environ. Microbiol.* **80**, 7021–7027 (2014).
175. Dolin, M. I. & Juni, E. Utilization of oxalacetate by Acinetobacter calcoaceticus: evidence for coupling between malic enzyme and malic dehydrogenase. *J. Bacteriol.* **133**, 786–793 (1978).
176. TAYLOR, W. H. & JUNI, E. Pathways for biosynthesis of a bacterial capsular

polysaccharide. I. Carbohydrate metabolism and terminal oxidation mechanisms of a capsuleproducing coccus. *J. Bacteriol.* **81**, 694–703 (1961).
177. Ni, B.-J. *et al.* CHAPTER 16 Denitrification Processes for Wastewater Treatment. in *Metalloenzymes in Denitrification: Applications and Environmental Impacts* 368–418 (The Royal Society of Chemistry, 2017). doi:10.1039/9781782623762-00368.
178. Lu, H., Chandran, K. & Stensel, D. Microbial ecology of denitrification in biological wastewater treatment. *Water Res.* **64**, 237–254 (2014).
179. Krehenbrink, M., Oppermann-Sanio, F.-B. & Steinbüchel, A. Evaluation of non-cyanobacterial genome sequences for occurrence of genes encoding proteins homologous to cyanophycin synthetase and cloning of an active cyanophycin synthetase from Acinetobacter sp. strain DSM 587. *Arch. Microbiol.* **177**, 371–380 (2002).
180. Elbahloul, Y. & Steinbüchel, A. Engineering the genotype of Acinetobacter sp. strain ADP1 to enhance biosynthesis of cyanophycin. *Appl. Environ. Microbiol.* **72**, 1410–1419 (2006).
181. Dickinson, K. E., Whitney, C. G. & McGinn, P. J. Nutrient remediation rates in municipal wastewater and their effect on biochemical composition of the microalga Scenedesmus sp. AMDD. *Algal Res.* **2**, 127–134 (2013).
182. Dickinson, K. E. *et al.* Simultaneous remediation of nutrients from liquid anaerobic digestate and municipal wastewater by the microalga Scenedesmus sp. AMDD grown in continuous chemostats. *J. Appl. Microbiol.* **118**, 75–83 (2015).
183. Praveen, P., Nguyen, D. T. T. & Loh, K. C. Biodegradation of phenol from saline wastewater using forward osmotic hollow fiber membrane bioreactor coupled chemostat. *Biochem. Eng. J.* **94**, 125–133 (2015).
184. Pieracci, J. P., Armando, J. W., Westoby, M. & Thommes, J. Chapter 9 - Industry Review of Cell Separation and Product Harvesting Methods. in (eds. Jagschies, G., Lindskog, E., Łącki, K. & Galliher, P. B. T.-B. P.) 165–206 (Elsevier, 2018). doi:https://doi.org/10.1016/B978-0-08-100623-8.00009-8.
185. Rodriguez, N. *et al.* The systems biology format converter. *BMC Bioinformatics* **17**, 154 (2016).
186. Seaver, S. M. D. *et al.* The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.* **49**, D575–D588 (2021).
187. Bryksin, A. V & Matsumura, I. Rational Design of a Plasmid Origin That Replicates Efficiently in Both Gram-Positive and Gram-Negative Bacteria. *PLoS One* **5**, e13244 (2010).

A. Appendix A. Chapter 2 Supplementary Information

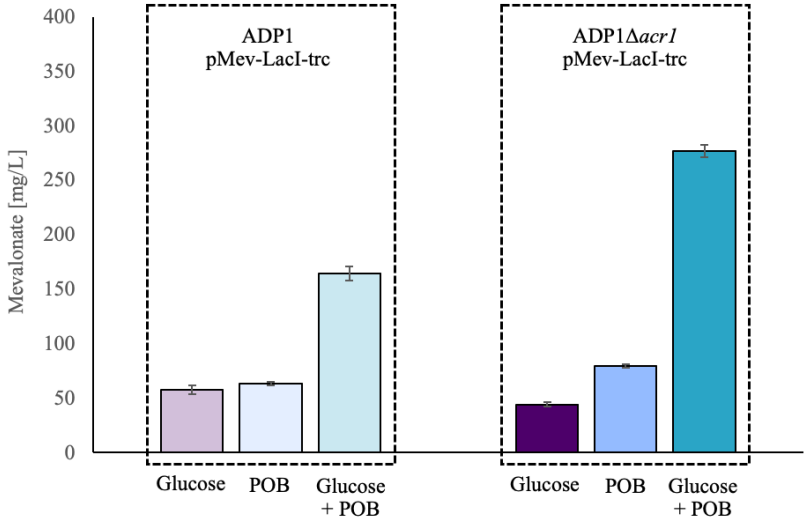

Figure A-1 Mevalonate is produced from a solely aromatic acid carbon source.

ADP1 pMev-LacI-trc (ECA10) and ADP1Δ*acr1* pMev-LacI-trc (ECA15) were cultivated in batch mode for 48 hours with either 10 mM glucose, or 20 mM POB, or 10 mM glucose and 20 mM POB as carbon sources. Data reflect mean and error bars are S.E.M. (n = 4).

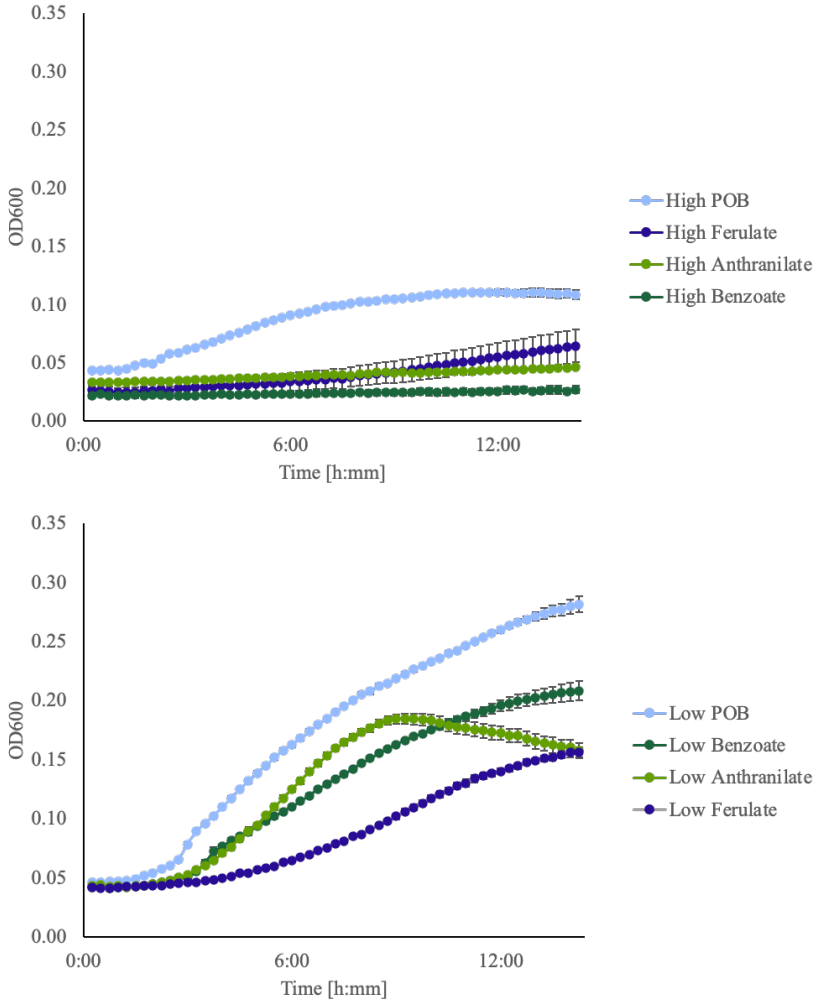

Figure A-2 ADP1 pMev-LacI-trc (ECA10) grows more quickly and to a higher OD in POB medium.

Growth of ADP1 pMev-LacI-trc (ECA10) was measured in a 96-well plate in medium containing 10 mM glucose with supplementation of (A) high (20 mM) or (B) low (5 mM) concentrations of POB, benzoate, anthranilate, and ferulate. Data reflect mean and error bars are S.E.M. (n = 4).

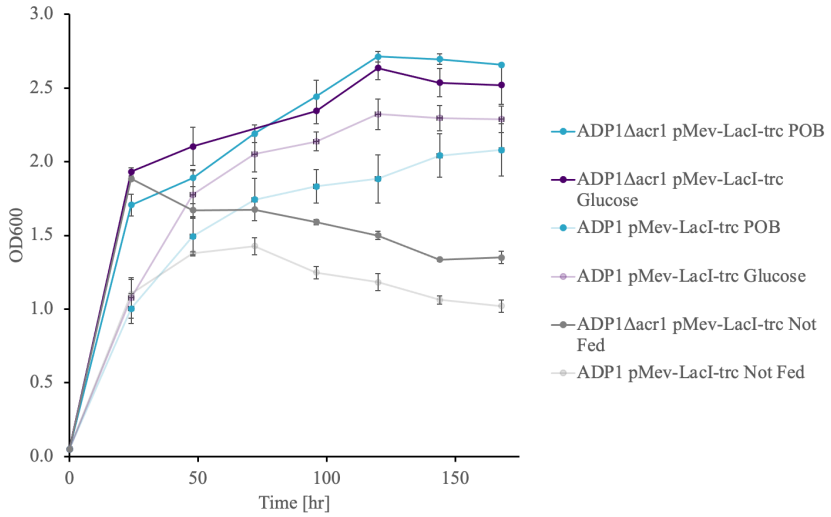

Figure A-3 Fed-batch cultures exhibit slowed growth after initial exponential phase.

OD600 was measured every 24 hours for fed-batch cultures of ADP1 pMev-LacI-trc and ADP1Δ*acr1* pMev-LacI-trc. Values are mean and error bars are S.E.M. (n = 6 for ADP1 pMev-LacI-trc POB, glucose, not fed, and for ADP1Δ*acr1* pMev-LacI-trc POB, and n = 2 for ADP1Δ*acr1* pMev-LacI-trc glucose and not fed).

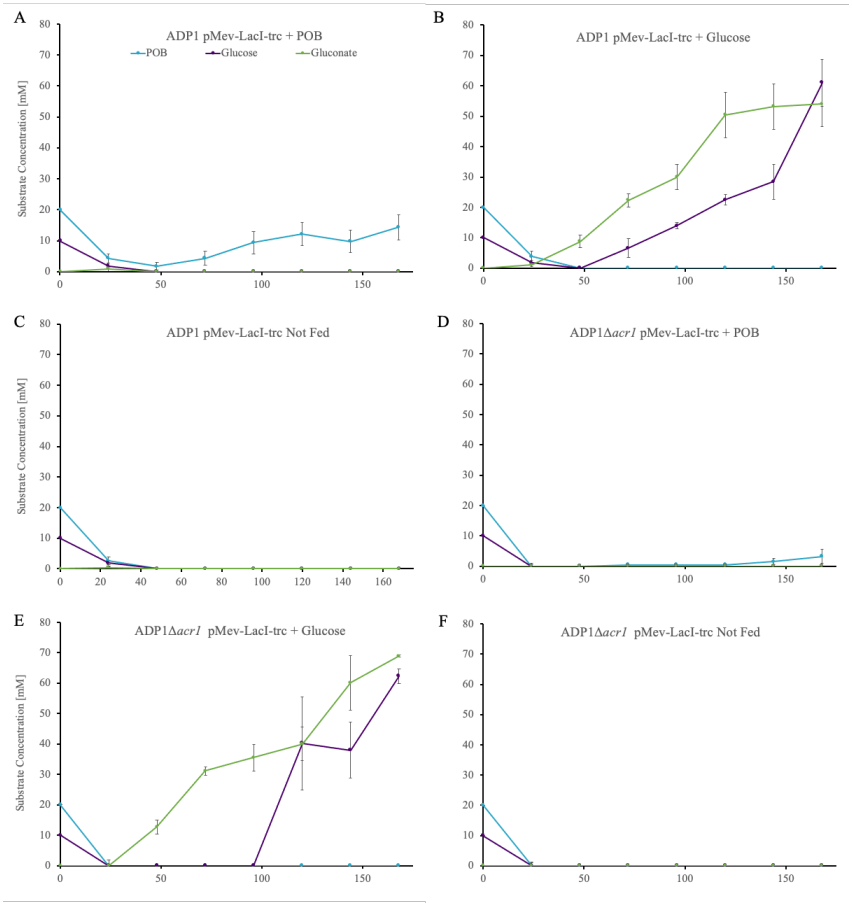

Figure A-4 Initial carbon sources are consumed by 48 hours.

Substrate concentrations (glucose and POB) and substrate-derived metabolite concentration (gluconate) were measured for fed-batch cultures. POB, glucose, and gluconate concentrations are plotted for ADP1 pMev-LacI-trc (A) + POB, (B) + glucose, and (C) not fed, and for ADP1Δ*acr1* pMev-LacI-trc (D) + POB, (E) + glucose, and (F) not fed. The vertical axis is substrate concentration, the horizontal axis is time in hours. Values are mean and error bars are S.E.M. (n = 6 for ADP1 pMev-LacI-trc POB, glucose, not fed, and for ADP1Δ*acr1* pMev-LacI-trc POB, and n = 2 for ADP1Δ*acr1* pMev-LacI-trc glucose and not fed).

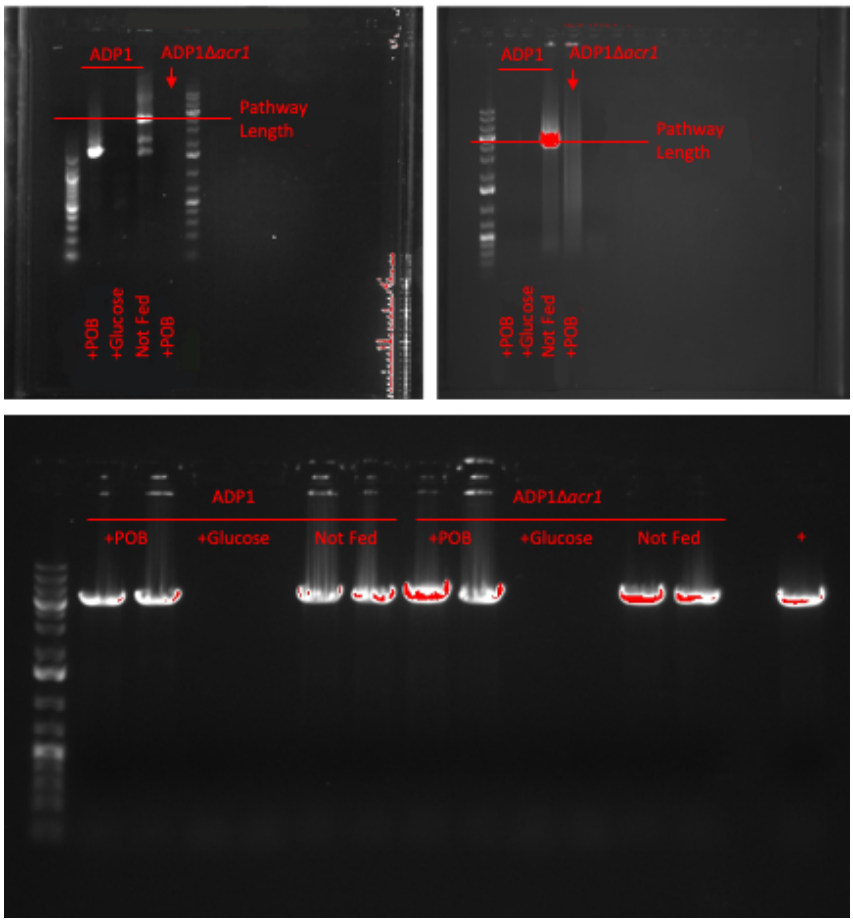

Figure A-5 Mutations to the mevalonate pathway primarily impact glucose-fed cultures.

Mevalonate pathway retention was assessed at 168 hours for fed-batch cultures. DNA gels show mutations to replicates with either no amplification of the pathway or amplification of a truncated pathway. POB fed cultures exhibited mutations for 4 of 12 cultures. Glucose fed cultures exhibited mutations for 6 of 8 cultures. Not fed cultures exhibited mutations for 1 of 8 cultures.

Table A.1 Primers used for the construction of strains in this study.

Primer	Primer Sequence	Binds to	Methods Section
ECA01	GGAATTGTGAGCGGATAACAATTTCAGAATTCAAAAGATCTTAGG	pJBEI-6410	Construction of plasmids
ECA02	TGACACTAGGTCTAGGGCGGGATAAAACGAAAGGCCCAGTCTTT	pJBEI-6410	Construction of plasmids
ECA03	ACTGGGCCTTTCGTTTTATCCCGCCCTAGACCTAGTGTC	pBWB162	Construction of plasmids
ECA04	CCTCCTAAGATCTTTTGAATTCTGAAATTGTTATCCGCTCACAAT	pBWB162	Construction of plasmids
ECA05	GCAAAAGTATGTCAAAGGAAAACCC	ADP1 Genome	*acr1* Knock Out
ECA06	AAAATCGTTTCTGAGACGTTTGTATTTGGATTGAAGACGGTTAAAGG	ADP1 Genome	*acr1* Knock Out
ECA07	ACCGTCTTCAATCCAAATACAAACGTCTCAGAAACGATTTTGAG	pBWB290	*acr1* Knock Out
ECA08	GGTCGTAACCATAAAAAAGCCAAAGGTTCTTGATGCTGAAACG	pBWB290	*acr1* Knock Out
ECA09	TTCAGCATCAAGAACCTTTGGCTTTTTTATGGTTACGACCATCAGCC	ADP1 Genome	*acr1* Knock Out
ECA10	GGGTGACGGCAGATGAAGG	ADP1 Genome	*acr1* Knock Out
ECA11	TACTTTTGGCGTGCAAGATGGTCG	ADP1 Genome	*acr1* Knock Out
ECA12	AAAAGCCTCTCGGTATGAGAGGATTGAAGACGGTTAAAGGGAAAT	ADP1 Genome	*acr1* Knock Out
ECA13	CCCTTTAACCGTCTTCAATCCTCTCATACCGAGAGGCTTTTTTATGG	ADP1 Genome	*acr1* Knock Out
ECA14	AACAGAAATGCTGTTTGATGTGGGC	ADP1 Genome	*acr1* Knock Out
ECA15	GAAACGATTGCCAAGATTGCCC	ADP1 Genome	*acr1* Knock Out
ECA16	GCTCACCGACCTTCTCATCG	ADP1 Genome	*acr1* Knock Out
ECA17	TACGCAAACCGCCTCTCC	pECA03	Culturing
ECA18	CCGCTATATAACACTTGATGAAAGCC	pECA04	Culturing
BWB645	AATTAACAGTTAACAAATAATAATACTAGTAGCGGCCGCT	pBWB162	Construction of plasmids
BWB646	GCAAACTTTTTGATGTTCATCTAGTATTTCTCCTCTTTCTCTAGT	pBWB162	Construction of plasmids
BWB647	AGAAAGAGGAGAAATACTAGATGAACATCAAAAAGTTTGCAA	sacB	Construction of plasmids
BWB648	AGCGGCCGCTACTAGTATTATTATTTGTTAACTGTTAATTGTCCTTG	sacB	Construction of plasmids

Table A.2 Bacterial strains used in this study.

Strain Name	Source	Description
Acinetobacter baylyi ADP1	Elliott & Neidle, 2011	Wild type *A. baylyi* ADP1
Escherichia coli DH5α	ThermoFisher	Wild type *E. coli* DH5α
ADP1 pMev-LacI-trc (ECA10)	This study	Wild type ADP1 harboring pECA03
ADP1Δ*acr1* (ECA14)	This study	Wax ester knockout strain, ADP1Δ*acr1*
ADP1Δ*acr1* pMev-LacI-trc (ECA15)	This study	Wax ester knockout strain harboring pECA03, pECA03 (ADP1Δ*acr1*)

Table A.3 Plasmids used in this study.

Plasmid Name	Source	Description
pJBEI-6410	Zhang, Nielsen, & Liu, 2017	Plasmid containing mevalonate pathway genes
pBWB162	Biggs et al., 2020	mCherry expressed under p_{trc} on the broad host-range vector, pBAV1k
pBWB290	This study	Constitutive Kan marker and SacB expressed under p_{trc} on pBAV1k vector
pMev-LacI-trc (pECA03)	This study	Mevalonate pathway genes, atoB, ERG13, and HMG1, expressed under p_{trc} on pBAV1k vector

Table A.4 Mutations in Fed-Batch Culture

Strain	Feed	Replicates with Mutations [No.]	Total Replicates [No.]
ADP1 pMev-LacI-trc	POB	2	6
ADP1 pMev-LacI-trc	Glucose	4	6
ADP1 pMev-LacI-trc	Not Fed	1	6
ADP1Δ*acr1* pMev-LacI-trc	POB	2	6
ADP1Δ*acr1* pMev-LacI-trc	Glucose	2	2
ADP1Δ*acr1* pMev-LacI-trc	Not Fed	0	2

Table A.5 168 Hour pH of Fed-Batch Cultures

Strain	Feed	168 Hour pH	Standard Error (n = 2)
ADP1 pMev-LacI-trc	POB	7.18	0.14
ADP1 pMev-LacI-trc	Glucose	4.27	0.13
ADP1 pMev-LacI-trc	Not Fed	7.51	0.01
ADP1Δ*acr1* pMev-LacI-trc	POB	9.37	0.04
ADP1Δ*acr1* pMev-LacI-trc	Glucose	4.31	0.02
ADP1Δ*acr1* pMev-LacI-trc	Not Fed	7.45	0.00

B. Appendix B. Chapter 3 Supplementary Information

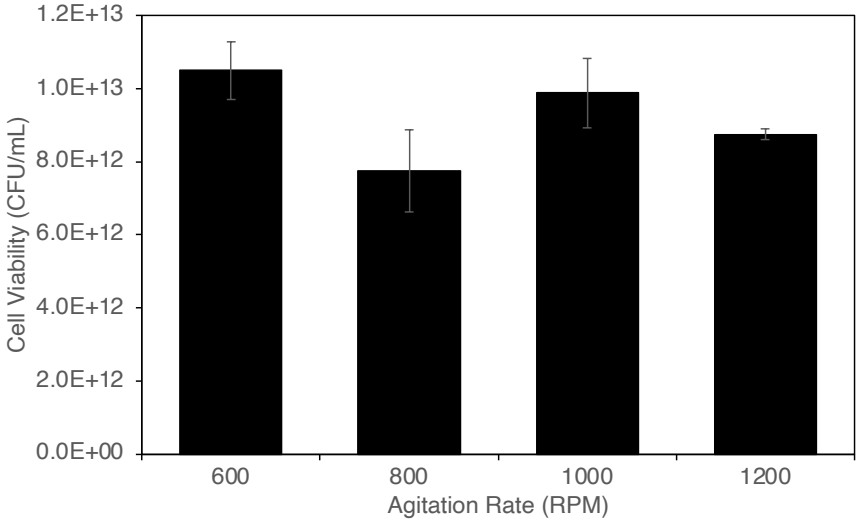

Figure B-1 ADP1 WT cell viability as colony forming units per mL (CFU/mL) at varied agitation rates from 600 – 1200 rpm in bioreactor cultures. N = 4 technical replicates. Error bars are S.E.M.

Table B-1 pH of initial media and endpoint samples from spent medium cultures estimated using pH paper.

pH				
Culture Number	Culture Name	Spent Media Timepoint (h)	Medium	48 hr
1.1	ADP1 Spent Medium + nutrients	24	7	7
1.2	ADP1 Spent Medium + nutrients	24	7	7
1.3	ADP1 Spent Medium + nutrients	24	7	7
1.4	ADP1 Spent Medium + nutrients	24	7	7
2.1	ADP1 Spent Medium + nutrients	72	7	7 - 8
2.2	ADP1 Spent Medium + nutrients	72	7	7 - 8
2.3	ADP1 Spent Medium + nutrients	72	7	7 - 8
2.4	ADP1 Spent Medium + nutrients	72	7	7 - 8
3.1	ADP1 Spent Medium + nutrients	120	7	7 - 8
3.2	ADP1 Spent Medium + nutrients	120	7	7 - 8
3.3	ADP1 Spent Medium + nutrients	120	7	7 - 8
3.4	ADP1 Spent Medium + nutrients	120	7	7 - 8
4.1	ADP1 Spent Medium	24	7	7 - 8
4.2	ADP1 Spent Medium	24	7	7 - 8
4.3	ADP1 Spent Medium	24	7	7 - 8
4.4	ADP1 Spent Medium	24	7	7 - 8
5.1	ADP1 Spent Medium	72	7	7 - 8
5.2	ADP1 Spent Medium	72	7	7 - 8
5.3	ADP1 Spent Medium	72	7	7 - 8
5.4	ADP1 Spent Medium	72	7	7 - 8
6.1	ADP1 Spent Medium	120	7	7 - 8
6.2	ADP1 Spent Medium	120	7	7 - 8
6.3	ADP1 Spent Medium	120	7	7 - 8
6.4	ADP1 Spent Medium	120	7	7 - 8
7.1	ADP1 Fresh Medium	--	7	7
7.2	ADP1 Fresh Medium	--	7	7
7.3	ADP1 Fresh Medium	--	7	7
7.4	ADP1 Fresh Medium	--	7	7 - 8
8.1	ADP1 Spent Medium + 1X M9	24	7	7 - 8
8.2	ADP1 Spent Medium + 1X M9	24	7	7 - 8
8.3	ADP1 Spent Medium + 1X M9	24	7	7 - 8
8.4	ADP1 Spent Medium + 1X M9	24	7	7 - 8
9.1	ADP1 Spent Medium + 1X M9	72	7	7 - 8
9.2	ADP1 Spent Medium + 1X M9	72	7	7 - 8
9.3	ADP1 Spent Medium + 1X M9	72	7	7 - 8
9.4	ADP1 Spent Medium + 1X M9	72	7	7 - 8
10.1	ADP1 Spent Medium + 1X M9	120	7	7 - 8
10.2	ADP1 Spent Medium + 1X M9	120	7	7 - 8
10.3	ADP1 Spent Medium + 1X M9	120	7	7 - 8
10.4	ADP1 Spent Medium + 1X M9	120	7	7 - 8

C. Appendix C. Chapter 4 Supplementary Information

C.1 Acinetobacter baylyi sp. ADP1 metabolic model preprocessing

The genome-scale metabolic model (GEM) for *Acinetobacter baylyi* sp. ADP1, titled "iAbaylyiv4", was downloaded from the BioModels database [119,120]. Because the SBML format of this model was the older L2V1 format and most tools and packages now require the L3V1 format, the model was converted to this newer format using the Systems Biology Format Converter (SBFC) [185]. Default flux bound constraints were then added using a custom python script (*add_flux_bounds.py*). Next, because many special characters were represented within model IDs as text (e.g. "[C]" was written as "_LBRACKET_C_RBRACKET_"), these IDs were cleaned up. Gene-Protein Rules (GPRs), which were previously stored in the "notes" section of the model for legacy purposes, were transferred to the GPR field of each reaction. Finally, each model metabolite was annotated with its ModelSEED ID, if available [186]. This updated and cleaner version of the iAbaylyiv4 model is provided in SBML format in the supplemental data.

C.2 Reaction rule filter – additional details

A custom python script, *map_rules.py*, was used to map rules to reactions by considering the cofactors involved as well as the bonds formed and broken (as well as any spectator atoms required for the rule). Because the reaction rules are written in the SMARTS chemical reaction language, we used the python package, RDKit, to determine whether a reaction could act on a given substrate during reaction-based mapping [123].

For gene-mapped rules, in order to determine an appropriate E-value cutoff for which rules to keep (as the default of 10^{-1} is far too lenient), we investigated the overlap of gene-mapped and reaction-mapped rules. Hits for overlapping rules (those both "gene-mapped" and "reaction-mapped") tended to be more significant (have a lower E-value) than those for nonoverlapping rules, validating the use of overlap as a metric of performance (Figure C.1).

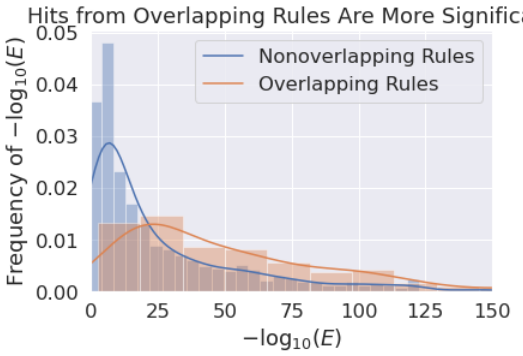

Figure C-1 Hits from overlapping (gene- and reaction-mapped) rules (red) are more significant (have lower E-values) than those that are only gene- or only reaction-mapped (blue).

Here, gene-mapped rules are those with $E < 10^{-1}$ (1 on x-axis).

We wanted to choose an E-value that maximizes this overlap while avoiding having too many gene-based rules that are not also reaction-mapped. We used the following equation to calculate an appropriate loss, L, as a function of E-value, E:

$$L(E) = n_g(E) - \alpha * n_{g,r}(E) \qquad \text{(C.1)}$$

where $n_g(E)$ is the number of gene-mapped rules sharing homology with at least one ADP1 gene at a specified E-value. $n_{g,r}(E)$ is the same but for rules that are both gene-mapped and reaction-mapped, representing the overlap between gene- and reaction-mapped rules at a given E-value. α is a scaling factor to account for the large difference in the number of gene-mapped reactions (up to 3,145 at $E=10^{-1}$) and reaction-mapped reactions that are also gene-mapped (up to 323 at $E=10^{-1}$). Specifically, α is the ratio of the maximum number of gene-mapped rules and overlapping rules (which occurs at an E-value threshold of 10^{-1}):

$$\alpha = \frac{n_g(E=10^{-1})}{n_{g,r}(E=10^{-1})} \qquad \text{(C.2)}$$

These equations were used to find the E-value at which the loss is minimized (after plotting $L(E)$ against E for $0 \leq -\log_{10} E \leq 300$) (Figure C.2). This optimal E-value was determined to be 10^{-35}, so only rules with at least one hit with $E \leq 10^{-35}$ were considered to be "gene-mapped". These "gene-mapped" rules were combined with the "reaction-mapped" rules to create the set of ADP1-specific rules.

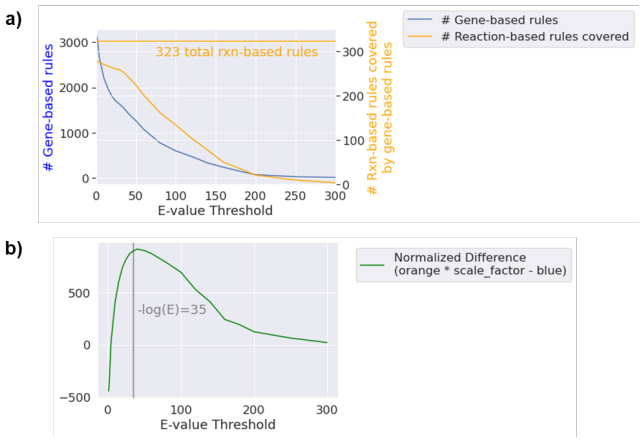

Figure C-2: E-value threshold is optimized to maximize overlap between gene-mapped and reaction-mapped rules (see Equations (1) and (2) in main text).

(a) As E-value threshold is made more stringent (x-axis is -log(E)), the total number of gene-based rules remaining (blue) decreases. The number of gene-based rules that also cover reactions (orange) also decreases, but with a different shape. (b) Using equations (1) and (2), the overlap between gene-based and reaction-based rules was used as a metric to find the optimal value ($E = 10^{-35}$) for gene-mapping.

C.3 Feature generation – additional details

Figure C.3 shows a summary of calculated features. Table C.1 lists all 26 features and their descriptions. See main text for an overview of how features were generated.

All features were calculated using a custom python script, *feature_set.py*. When calculating fquin_dist_from_feed and fsucc_dist_from_feed cofactors were excluded when finding shortest

paths through the network. Feed rate was set to 20 mmol/gDCW/hr for either quinate or succinate when calculating maximum theoretical yield using the ADP1 genome-scale model.

Because not all features were able to be calculated for every candidate (e.g. deepRFC did not accept the SMILES for some candidates), a value for that feature was either assigned or sampled from that feature column. Specifically, log_fold_change was assigned 1; overall_peaks_frac, peaks_frac_ratio, quinate_peaks_frac, and succinate_peaks_frac were assigned 0; and fquin_dist_from_feed, fsucc_dist_from_feed, dff_fc_diff, dff_fc_prod, dist_from_feed_log_ratio, feasible, feasible_frac, rt_dev_diff, and rt_dev_pct were sampled from that respective feature column. All sampling was done with a preset seed of 42.

Most features which could not be calculated were due to divide by zero errors (e.g. peaks_frac_ratio). The only other major issue was orphan reactions in the metabolic model which occasionally prevented finding any path in the metabolic network from a feed metabolite to a candidate compound to calculate fquin_dist_from_feed and fsucc_dist_from_feed. This occurred for roughly 20% of candidate-peak pairs, primarily for candidates produced from large secondary metabolites which are not connected to quinate or succinate in the metabolic model. All other features requiring some sampling (e.g., rt_dev_diff) had less than 5% candidate-peak pairs that had to be sampled.

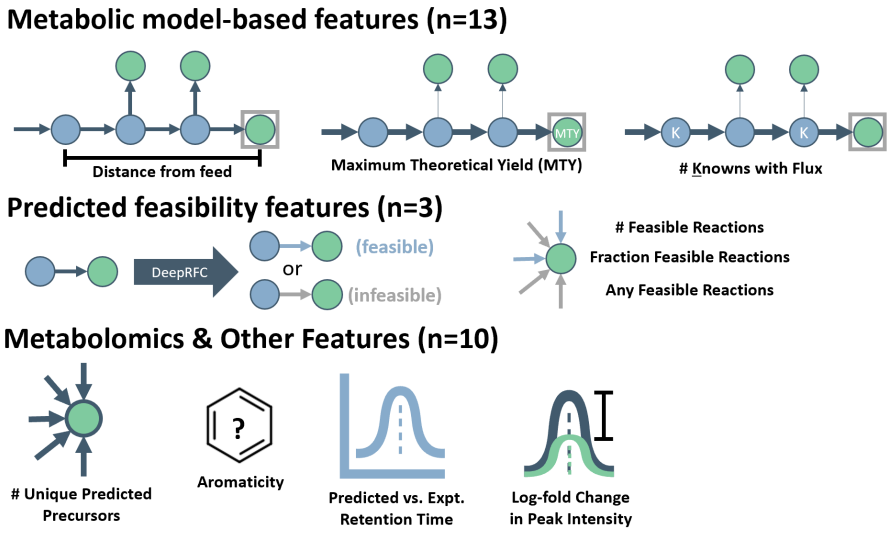

Figure C-3 The ADP1 metabolic model, DeepRFC, as well as other data sources and tools were used to calculate features of each candidate to be used in machine learning classification

Calculated features were categorized into 3 separate sets. (a) 13 features were calculated by utilizing the APD1 extended metabolic model, including the distance between a candidate and the feed metabolite in the metabolic network, the predicted maximum theoretical yield of a candidate, and the number of experimentally validated (from the metabolomics dataset) compounds predicted to hold flux [119]. (b) deepRFC was used to classify predicted reactions as feasible or infeasible [128]. Because many candidates are the product of more than one predicted reaction, the number of feasible reactions as well as fraction of all predicted reactions that were classified as feasible were used as features. (c) Features calculated from metabolomics data (e.g., from retention time, peak intensities) were also used. In addition, some other features that don't fit in the first two categories (e.g., aromaticity, number of unique precursors) were added to this category.

Table C.1 Description of all 26 features used in the final model

#	Category	Name	Description
1	Metabolic Modeling	fquin_dist_from_feed	Shortest distance in # reactions separating candidate and quinate (feed) in metabolic model
2	Metabolic Modeling	fsucc_dist_from_feed	Shortest distance in # reactions separating candidate and succinate (feed) in metabolic model
3	Metabolic Modeling	dist_from_feed_log_ratio	$\log_{10}$(fquin_dist_from_feed / fsucc_dist_from_feed). How close a candidate is to quinate vs succinate in the metabolic network.
4	Metabolic Modeling	dff_fc_diff	dist_from_feed_log_ratio - log_fold_change. High if close to quinate in network and candidate peak is more abundant in quinate feeding condition, low if close to succinate in network and candidate peak is more abundant in succinate feeding condition.
5	Metabolic Modeling	dff_fc_prod	dist_from_feed_log_ratio * log_fold_change. Low if close to given feed (quinate or succinate) AND candidate peak is more abundant with that feed.
6	Metabolic Modeling	g0_fquin_mty	Predicted maximum theoretical yield (mty) of candidate with no growth requirement (g0) with quinate feed
7	Metabolic Modeling	g0_fsucc_mty	Predicted maximum theoretical yield (mty) of candidate with no growth requirement (g0) with succinate feed
8	Metabolic Modeling	g0_fquin_mty_mass	g0_fquin_mty / (candidate mass). MTY when fed quinate on a mass rather than molar basis.
9	Metabolic Modeling	g0_fsucc_mty_mass	g0_fsucc_mty / (candidate mass) . MTY when fed succinate on a mass rather than molar basis.
10	Metabolic Modeling	g0_fquin_n_knowns_w_flux	# of experimentally validated (known) compounds involved in flux-carrying reactions in the flux distribution from the g0_fquin_mty calculation
11	Metabolic Modeling	g0_fsucc_n_knowns_w_flux	# of experimentally validated (known) compounds involved in flux-carrying reactions in the flux distribution from the g0_fsucc_mty calculation
12	Metabolic Modeling	g0_fquin_norm_n_knowns_w_flux	g0_fquin_n_knowns_w_flux / (# all metabolites involved in flux-carrying reactions during g0_fquin_mty calculation), i.e. the fraction of flux-carrying metabolites from MTY calculation that are experimentally validated.
13	Metabolic Modeling	g0_fsucc_norm_n_knowns_w_flux	g0_fquin_n_knowns_w_flux / (# all metabolites involved in flux-carrying reactions during g0_fsucc_mty_calculation), i.e. the fraction of flux-carrying metabolites from MTY calculation that are experimentally validated.

14	Predicted Feasibility	feasible	Whether a candidate has any predicted reactions predicted to be feasible via DeepRFC
15	Predicted Feasibility	feasible_frac	Fraction of all predicted reactions producing candidate predicted to be feasible via DeepRFC
16	Predicted Feasibility	feasible_n	# predicted reactions producing candidate predicted to be feasible via DeepRFC
17	Metabolomics & Others	rt_dev_diff	Difference between predicted and experimental retention time values. See Appendix B.5 for retention time prediction methods.
18	Metabolomics & Others	rt_dev_pct	Percent difference between predicted and experimental retention time values. See Appendix C.5 for retention time prediction methods.
19	Metabolomics & Others	log_fold_change	Log fold change in peak intensity for quinate vs succinate feed conditions (values taken straight from the metabolomics dataset). Min and max set to -5 and 5 (due to 0, inf values).
20	Metabolomics & Others	overall_peaks_frac	Proportion of all samples in which the peak for this peak-candidate pair was detected (values taken straight from the metabolomics dataset).
21	Metabolomics & Others	quinate_peaks_frac	Proportion of quinate-fed samples in which the peak for this peak-candidate pair was detected (values taken straight from the metabolomics dataset).
22	Metabolomics & Others	succinate_peaks_frac	Proportion of succinate-fed samples in which the peak for this peak-candidate pair was detected (values taken straight from the metabolomics dataset).
23	Metabolomics & Others	peaks_frac_log_ratio	$\log_{10}$(quinate_peaks_frac / succinate_peaks_frac). High if candidate peak detected more often when fed quinate, low if candidate peak detected more often when fed succinate.
24	Metabolomics & Others	n_known_precursors	# experimentally verified (known) precursor compounds for candidate. Metabolism is highly interconnected, so novel compounds are likely made via multiple reactions as well.
25	Metabolomics & Others	n_unique_precursors	# unique precursor compounds for candidate. Metabolism is highly interconnected, so novel compounds are likely made via multiple reactions as well.
26	Metabolomics & Others	aromatic	Whether a candidate contains any aromatic rings. May be useful because quinate is a fed aromatic compound while succinate is not.

C.4 Classifier optimization and feature importances

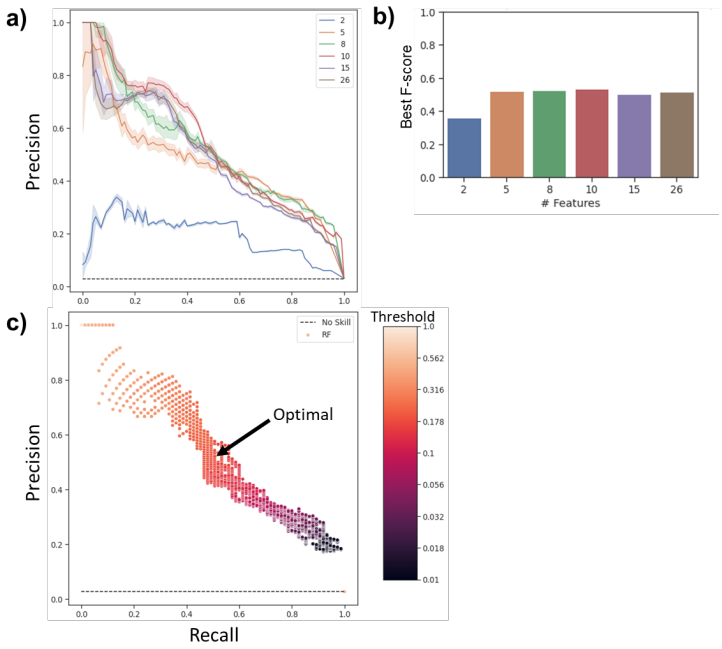

Figure C-4 Feature reduction down to the top 10 features at a threshold of 0.30 maximized classifier performance on out-of-bag samples

(a) Precision-Recall (PR) curves after selecting the top *n* features for *n* = 2, 5, 8, 10, 15, and 26 (all features) across 10 random states. PR curves are generated by varying classifier threshold from 0 to 1. (b) The best F-score from the associated PR curve in (a) at the optimal threshold for that curve. This shows that performance is maximized with 10 features and doesn't significantly decrease until less than 5 features are used. (c) PR curve for *n* = 10 where metrics are plotted individually for each threshold tested and each random state used (equivalent to red line in (a) but now with threshold highlighted as shown by colorbar). Optimal performance was at threshold = 0.30, designated by the black arrow (F = 0.50).

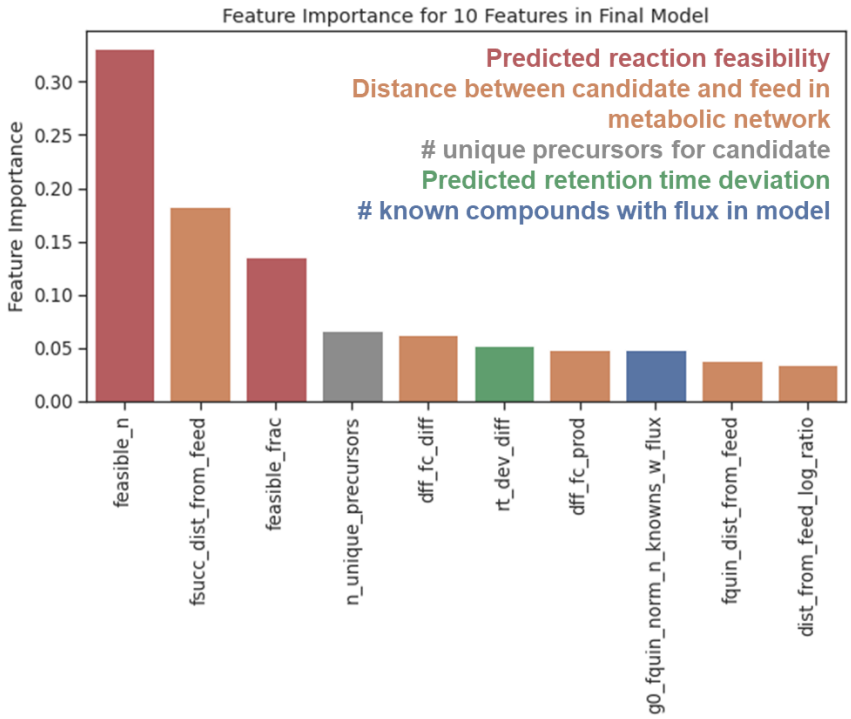

Figure C-5 The top 10 features used in the final classifier include features from all three feature sets and can be divided into 5 sub-categories (differentiated by color)

Predicted reaction feasibility-based features (red) were found to be most important. Modeling-based features (orange and blue) were also found to be highly important. Metabolomics and other features (grey and green) were useful, albeit less so than the other two feature sets.

C.5 Retention time prediction

In order to maximize the input of the metabolomics dataset in our classification workflow, we wanted to use not only the m/z data, but also the retention times associated with each peak. However, we do not know the retention times of predicted candidates, as it is not a straightforward calculation (unlike calculation of exact mass to compare m/z). Predicting retention time from

chemical structure is notoriously difficult, as it depends not only on the chemical structure but also the column and chromatographic conditions used. To control for these confounding variables, we built a QSAR (quantitative structure-activity relationship) model from the 105 experimentally validated compounds in the ADP1 metabolomics dataset used (where all compounds were processed in the same chromatography step). These 105 compounds' retention times are plotted in Figure B.6 (blue). Retention times for unknown peaks are also plotted (red) and are similarly distributed.

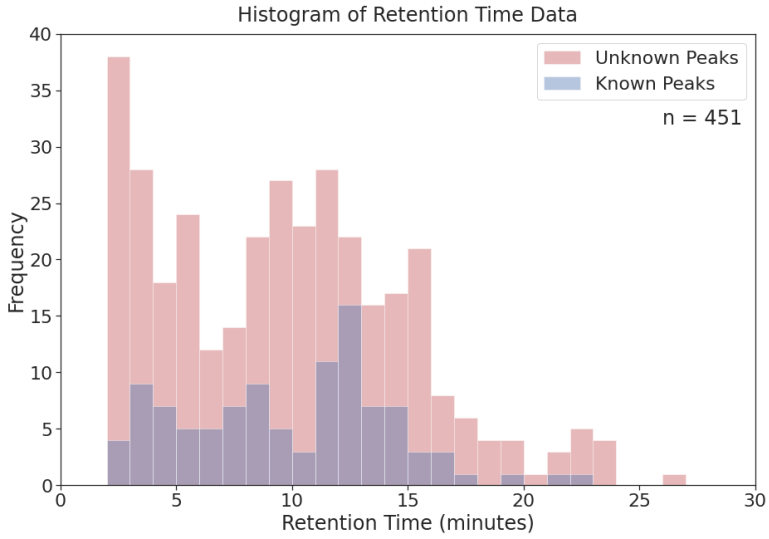

Figure C-6 Retention times of known peaks have a similar distribution to that of unknown peaks in the ADP1 metabolomics dataset

Retention times are graphed separately for known peaks (blue, n=105) and unknown peaks (red, n=346).

We used scikit-learn to build a random forest regressor to predict retention time from structure, trained on these 105 compounds [130]. We featurized each structure by using mordred, a python library that calculates 1,613 2D and 213 3D features for a given compound (e.g. number of acid

sites) [139]. We then performed a feature reduction step (reducing down to the top 50 features) and assessed performance based on out-of-bag prediction. We were able to achieve an R^2 of 0.64 on out-of-bag samples with a mean error of 2.0 minutes (Figure C.7).

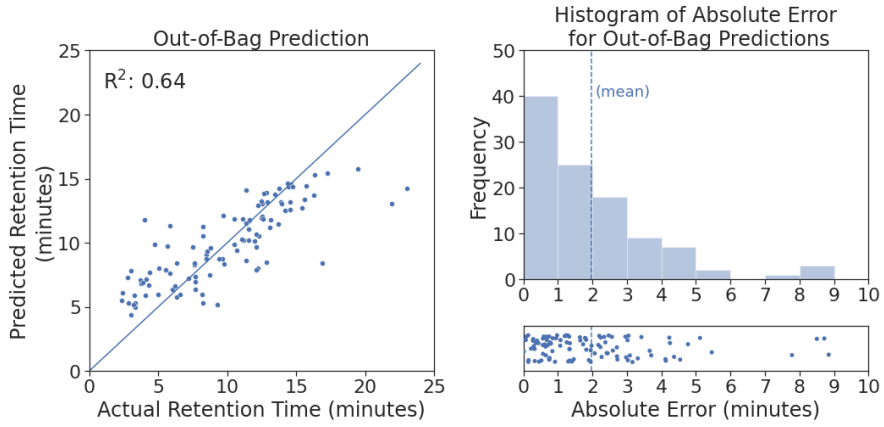

Figure C-7 A mean error of 2.0 minutes is attained by training a random forest regressor to predict compound retention time from structure

(left) Prediction of retention time on out of bag samples. (right) Histogram of error for each out-of-bag sample, with individual samples shown below.

We judged that these predictions were not high-resolution enough to be used as a binary filter to filter out candidates (95% of out-of-bag predictions were within 5.2 min which we considered too much deviation from the measured value). However, we still wanted to try and use these predicted values if possible. Thus, we included retention time predictions when calculating features for each candidate (rt_dev_diff and rt_dev_pct). This allowed us to (1) let the learner tell us if these retention time predictions were at all useful for classification of compounds for known peaks and (2) if useful, utilize these predictions to help classify candidates for unknown peaks.

Table C.2 Primers used in cloning of knock-out cassettes for experimental validation

Primer Name	Sequence	Knock-out	Binds To
ECA158	tcgatgggcattcatattcaaccc	ACIAD1826	ADP1 Genome
ECA159	tttttattggtgagaatccagataatttcctatgttgctgaaacgattgattg	ACIAD1826	ADP1 Genome
ECA160	cagcaacataggaaattatctggattctcaccaataaaaaacgccc	ACIAD1826	pBWB162
ECA161	atgagtttcacatttatgaaaaattctatcataattgtggtttcaaaatcggc	ACIAD1826	pBWB162
ECA162	ccacaattatgatagaatttttcataaatgtgaaactcatgtgcaactcacg	ACIAD1826	ADP1 Genome
ECA163	cctcaaccacgcaatcaaatgc	ACIAD1826	ADP1 Genome
ECA146	acaatctcccaaccttgcgg	ACIAD0381	ADP1 Genome
ECA147	tttttattggtgagaatccagatcaagatatcgctgaagtcgatgaac	ACIAD0381	ADP1 Genome
ECA148	acttcagcgatatcttgatctggattctcaccaataaaaaacgccc	ACIAD0381	pBWB162
ECA149	aatggatcatgaggtgaagcggacgtcaaattctatcataattgtggtttcaaaatcg	ACIAD0381	pBWB162
ECA150	ttatgatagaatttgacgtccgcttcacctcatgatccattatttaacg	ACIAD0381	ADP1 Genome
ECA151	gtcgtactgttggaccgatgg	ACIAD0381	ADP1 Genome
ECA152	atgcgactgtgattggtcagc	ACIAD0984	ADP1 Genome
ECA153	cgtttttttattggtgagaataaggcaagtcaatctttttttgattgagc	ACIAD0984	ADP1 Genome
ECA154	aaaaaagattgacttgccttattctcaccaataaaaaacgcccgg	ACIAD0984	pECA19
ECA155	tttacttaaacggcatggttattctatcataattgtggtttcaaaatcggc	ACIAD0984	pECA19
ECA156	aaccacaattatgatagaataaccatgccgtttaagtaaaaaactaaagattaagc	ACIAD0984	ADP1 Genome
ECA157	ggagaggtagataaccgtcatttaagtgg	ACIAD0984	ADP1 Genome
ECA188	gcattcttttgcttatggcccc	ACIAD3540	ADP1 Genome
ECA189	tttttattggtgagaatccatagttgattaaaaaaatcaaggtcaatccag	ACIAD3540	ADP1 Genome
ECA190	ttgatttttttaatcaactatggattctcaccaataaaaaacgcc	ACIAD3540	pBWB162
ECA191	aattaattataaaattgtaggacgtcaaattctatcataattgtggtttc	ACIAD3540	pBWB162

ECA192	ttatgatagaatttgacgtcctacaattttataattaattaagctaaatccatgccac	ACIAD3540	ADP1 Genome
ECA193	tggtatcgatatctacagctttaatcttgc	ACIAD3540	ADP1 Genome

Table C.3 Plasmids used in cloning of knock-out cassettes for experimental validation

Plasmid Name	Source
pBWB162 (pBAV1k-LacI-trc-mCherry)	Bryskin *et al.* [187]
pECA19 (pBAV1k-LacI-trc-mCherry (CmR))	This study

Vita

Erika Arvay was born and raised in Columbia, South Carolina. She graduated from Spring Valley High School in 2013 before attending Clemson University from 2013 – 2017 to earn her B.S. Chemical and Biomolecular Engineering with an emphasis area of Biomolecular Engineering. While at Clemson, Erika conducted research focused on metabolic engineering under the guidance of Dr. Mark Blenner.

After graduation, Erika attended Northwestern University to obtain a Ph.D. in Chemical and Biological Engineering where she worked under the guidance of Dr. Keith Tyo. She was awarded a cluster fellowship as well as a traineeship in the Biotechnology Training Program. After completion of her Ph.D she plans to take on the role of Scientist in the field of biotechnology.