Spring 1-1-2018

Application of Advances in Synthetic Biology to Modular Megasynthase Design and Scalable Combinatorial Biosynthesis

William Clifford Grau
University of Colorado at Boulder, bgrau03@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/chem_gradetds
Part of the Biomedical Engineering and Bioengineering Commons, Chemical Engineering Commons, and the Chemistry Commons

Recommended Citation
https://scholar.colorado.edu/chem_gradetds/252

This Dissertation is brought to you for free and open access by Chemistry & Biochemistry at CU Scholar. It has been accepted for inclusion in Chemistry & Biochemistry Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Application of advances in synthetic biology to modular megasynthase design and scalable combinatorial biosynthesis

by

William Clifford Grau

B.S., California Polytechnic State University, San Luis Obispo, 2012

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Chemistry and Biochemistry
2018
This thesis entitled:

Application of advances in synthetic biology to modular megasynthase design and scalable combinatorial biosynthesis

written by William Clifford Grau
has been approved for the Department of Chemistry and Biochemistry

________________________

Dr. Maciej Walczak

________________________

Dr. Ryan T. Gill

Date _______________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above-mentioned discipline.
Grau, William Clifford (Ph.D., Chemistry and Biochemistry)
Application of advances in synthetic biology to modular megasynthase design and scalable combinatorial biosynthesis
Thesis directed by Professor Ryan T. Gill

Abstract.

The diversity in structure and biological activity of natural products has long captured the imagination of chemists and molecular biologists alike. Natural products are synthesized by elaborate secondary metabolic pathways or enzymes. For example, modular megasynthases such as Type I Polyketide Synthases (PKSs), Non-ribosomal Peptide Synthetases (NRPSs), and PKS-NRPS hybrids, produce an array of quintessential natural products. The apparent modularity, defined architecture, and predictable chemistry of modular megasynthases make them especially attractive for combinatorial biosynthesis, which has long been a focus of biotechnology. Combinatorial biosynthetic libraries are of particular interest due to the high value of many natural products and for the potential directed evolution of small molecules.

Prior efforts to rationally design modular megasynthases have demonstrated some success but combinatorial forward engineering remains unfulfilled. The rise of synthetic biology has presented new capabilities to resolve the underlying deficiencies of past approaches. Synthetic biology abstracts complex biological systems into modular parts with simplified rules governing functions and interactions, enabling forward engineering of novel biological systems.

Herein, we present a precise and scalable synthetic biology approach to combinatorial biosynthesis built on improved abstraction of modular megasynthases into ‘parts’. The approach begins with bottom-up design of a chimeric modular megasynthase chassis using conserved catalytic domain linkers as modular parts for connecting catalytic domains. This minimal modular megasynthase ‘chassis’ is then assembled to produce a target molecule and finally diversified at scale to produce a library of new molecules using paralogous catalytic domains.

To demonstrate this approach, we first targeted the production of a triketide lactone food and fragrance ingredient, d-hexalactone. A retrosynthetic analysis identified catalytic domains, linkers, and substrates required for d-hexalactone biosynthesis via a PKS. The chimeric megasynthase chassis was then designed using known catalytic domains and a custom computational pipeline that identified context-independent linker sequences, built, and tested. To demonstrate the extensibility of this approach we designed, built, and tested Non-Ribosomal Peptide Synthetase (NRPS) and NRPS-PKS chassis. Finally, we applied a massively multiplex genome engineering tool for further rational diversification, providing a scalable, high efficiency approach to combinatorial biosynthesis.
Acknowledgements.

As I have begun writing this thesis, in fits and starts, it has occurred to me how retrospective the process is or more aptly, how it is strangely both retrospective and prospective. I have found myself reading memoirs and autobiographies in my spare time: Blood, Bones, and Butter by Gabrielle Hamilton, Speak, Memory by Vladimir Nabokov, My Struggle by Karl Ove Knausgaard. When my writing begins to slow, my mind accelerates. I consider how exactly I came to graduate school, what I sought to achieve, what my experience has been, whether or not I have achieved my goals, what my future holds, and what role my graduate experience will play in that future. Of course, I also wonder whether or not I have learned and grown enough to be deserving of a doctorate.

When I started graduate school, the notion that it would be a journey seemed like an unoriginal, tired cliché and yet, now that I’m done, it is of course the absolute most appropriate description. It turns out that my experience did not have to be particularly unique to be deeply important. To this end, it occurs to me that elaborating on this journey may at the very least help me sift through everything that is on my mind.

My path to graduate school began during my first year of undergrad. Midway through that year I was in and out of the hospital. Unsurprisingly, this completely shaped my decisions and worldview. Over the next few years, I completed degrees in Biochemistry and Finance, did a finance internship at Amgen, worked in a lab, and all the other things a college student can and should do. I left these experiences with an overwhelming sense of the power of combining great science with great business. A specific medicine allowed me to have each of these experiences and it struck me that this medicine was a perfect embodiment of this concept.

With that, I set out to learn as much as I could about both fields, and all the advice I received was to go to graduate school. My worldview made me focus entirely on identifying a graduate project that could have commercial viability. This, as it turns out, was the height of arrogance, and over the long run, significantly harmed my graduate experience.

The problem with this approach to graduate school was that I initially focused on generating commercializable IP as fast as I could, rather than on learning how to be the best scientist possible. This is why the approach was so arrogant, I didn’t fully understand how difficult it is to be a functional and useful scientist. As a result, I picked a project that was broad and risky. This had all kinds of implications - working alone outside the expertise of our lab, relying on unproven techniques and technologies, etc. Like any graduate student, I experienced a lot of failures. I was also on an island, with little support. I learned a lot, and these lessons were undeniably more painful because I saw them as a barrier, rather than an opportunity to learn. In the long run, the failures instilled some hard earned humility. Perhaps feedback along the way could have steered me onto the right path but, as I didn’t receive any and probably didn’t ask for enough, that is unknowable now.

My approach to these failures and lessons was devoid of finesse and nuance, I plowed through them with sheer force of will. Perhaps this mindset was a side effect of a ‘move fast and break things’ culture. What I am now realizing is how much energy this approach required. Working as much as humanly possible for months at a time is sapping. I used the hope of career success as the fuel for these efforts: if I was able to just get the IP I needed to start that company, it would all be worth it or, worst case scenario I would have earned the respect of my advisor, who would help me find a soft landing spot. Neither of these outcomes materialized and I was completely burnt out.
Luckily, my life has changed significantly since this all started. Namely, I married the love of my life. Our wedding took the most important lesson of graduate school, which had been simmering on the back burner of my mind and brought it to a boil. I burned out because my motivation was superficial. Career success can be an effective motivator but, if it doesn’t happen when or how one imagines, that fuel runs out and it becomes near impossible to keep moving. Getting married refueled me. As I think back, it is the outpouring of love from all of our friends and family that reenergized me.

Now my perspective on graduate school and on life in general is completely different. Is a successful career important? Of course, but far more important are my relationships. Using career success as a marker of value makes my value contingent on all kinds of exigencies. Our wedding made me realize that having strong relationships made me feel intrinsically valuable, and that this is a much more constant and stable source to draw energy from.

Through this lens it became clear that the only way I made it through graduate school was thanks to relationships; with the postdocs who taught me basically everything I know about biological engineering (Ramsey and Andrew); with Marcelo, who also patiently taught me so much (after I finished ranting over coffee, lunch, beers, on the bus, in the lab, or over text); with everyone whom I crossed paths with in the lab that made it such an incredibly supportive and enriching environment; with my friends who were always ready to kick around a soccer ball after work (Steve and Dylan) or who were ready to talk through things or distract me (Mikey, Joey, Nick, Nate, David, Brad, Matt, Sam, and Brent); with my sister, who always provided a place to crash and someone to adventure with when I needed to get away; with my Mom, who is undeniably the most patient listener and is full of completely boundless unconditional love; with my Dad, who is also full of unconditional love but in a different form, as a steadying force.

Of course, the most important person is last. Erin, my wife, saw me through this. I am stunned how she was and is always able to be exactly the person I need, at exactly the right moment: whether acting like total goofballs, having a deep talk about my existential crisis du jour, adventuring, or just hanging out. The amount of giving and selflessness that that requires is beyond me. Ultimately, the small semblance of perspective on life that I do have is thanks to her. Our wedding, that was so clarifying for me? All her. In fact, here is what was on my mind the morning of our wedding:

“You know it is difficult for me to translate my feelings into words in these situations. I have searched for but never found that single word or single sentence that adequately conveys the depth of my feelings for you. But, over the past few months, one word keeps emerging out of all these feelings I can’t put words to.

I am blessed to stand up here today and marry the most stunning person I have ever met. I am blessed every day to be a witness as you bring your love and passion to the world with such incredible force, energy, and grace. I am, we are, blessed with unconditional love, from each other, from our parents, from family, and from friends.

And so, I promise to be your husband and partner as we build a relationship that honors all the blessings we have received. I promise to be anything and everything you need me to be; your best friend who you can be a goofball with when no one (or everyone) is watching, or your lover who never wavers. I love you so much and I can’t wait for the rest of our lives together.”
I can’t imagine having had to do this without her, and I am overjoyed I get to be her partner through life’s trials and tribulations. I will consider myself a success if I can be for others and most importantly, for her, even a tiny portion of what she is for me.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Complex Adaptive Systems</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Evolution</td>
<td>12</td>
</tr>
<tr>
<td>1.3 Evolution as a biological engineering paradigm</td>
<td>17</td>
</tr>
<tr>
<td>1.4 Application of directed evolution to drug discovery</td>
<td>20</td>
</tr>
<tr>
<td>1.5 Directed evolution of small molecules in chemical biology and drug discovery</td>
<td>22</td>
</tr>
<tr>
<td>1.6 Modular megasynthases, combinatorial biosynthesis, and directed evolution</td>
<td>28</td>
</tr>
<tr>
<td>1.7 This work</td>
<td>28</td>
</tr>
<tr>
<td>1.8 References</td>
<td>29</td>
</tr>
<tr>
<td><strong>2. Computational Part Mining for Modular Megasynthase Engineering</strong></td>
<td>39</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>39</td>
</tr>
<tr>
<td>2.2 Results</td>
<td>45</td>
</tr>
<tr>
<td>2.3 Discussion</td>
<td>55</td>
</tr>
<tr>
<td>2.4 Materials and Methods</td>
<td>57</td>
</tr>
<tr>
<td>2.5 References</td>
<td>60</td>
</tr>
<tr>
<td><strong>3. Bottom-up, <em>de novo</em> design, build, and test of a novel modular megasynthase in <em>E. coli</em></strong></td>
<td>64</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>64</td>
</tr>
<tr>
<td>3.2 Results</td>
<td>70</td>
</tr>
<tr>
<td>3.3 Discussion</td>
<td>110</td>
</tr>
<tr>
<td>3.4 Materials and Methods</td>
<td>112</td>
</tr>
<tr>
<td>3.5 References</td>
<td>128</td>
</tr>
<tr>
<td><strong>4. Design extensibility and scaling with massively multiplexed genome engineering</strong></td>
<td>134</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>134</td>
</tr>
<tr>
<td>4.2 Results</td>
<td>136</td>
</tr>
<tr>
<td>4.3 Discussion</td>
<td>149</td>
</tr>
<tr>
<td>4.4 Materials and Methods</td>
<td>151</td>
</tr>
<tr>
<td>4.5 References</td>
<td>153</td>
</tr>
<tr>
<td><strong>5. Conclusions and future work</strong></td>
<td>156</td>
</tr>
<tr>
<td>5.1 Conclusions</td>
<td>156</td>
</tr>
<tr>
<td>5.2 Future work</td>
<td>157</td>
</tr>
<tr>
<td><strong>Bibliography</strong></td>
<td>159</td>
</tr>
</tbody>
</table>
# Table of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Target molecule chemical properties</td>
<td>139</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>1.1 The yeast protein-protein interaction network</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Darwin’s finches</td>
<td>7</td>
</tr>
<tr>
<td>1.3 Canonical approach to directed evolution</td>
<td>17</td>
</tr>
<tr>
<td>1.4 Polyketide structures</td>
<td>23</td>
</tr>
<tr>
<td>1.5 Modular megasynthese architecture</td>
<td>24</td>
</tr>
<tr>
<td>1.6 Structure of a PKS extension module</td>
<td>26</td>
</tr>
<tr>
<td>2.1 Computational workflow for identifying context-independent linkers</td>
<td>44</td>
</tr>
<tr>
<td>2.2 Computational workflow for linker mining and design and complete gene design</td>
<td>47</td>
</tr>
<tr>
<td>2.3 Phylogenetic tree of Ketosynthese-Acyltransferase linkers</td>
<td>48</td>
</tr>
<tr>
<td>2.4 Histogram of Ketosynthese-Acyltransferase (KS-AT) linker lengths</td>
<td>49</td>
</tr>
<tr>
<td>2.5 Whitaker plot of KS-AT linker sequence enrichments</td>
<td>50</td>
</tr>
<tr>
<td>2.6 Whitaker plot of KS-AT linker sequence enrichments</td>
<td>52</td>
</tr>
<tr>
<td>2.7 Phylogenetic analysis for identifying a single design</td>
<td>53</td>
</tr>
<tr>
<td>2.8 Comparison of predicted KS-AT linker to structure of a WT KS-AT linker</td>
<td>55</td>
</tr>
<tr>
<td>3.1 Molecular biology workflow to produce a target strain</td>
<td>70</td>
</tr>
<tr>
<td>3.2 Retrosynthetic analysis of δ-hexalactone</td>
<td>71</td>
</tr>
<tr>
<td>3.3 Modules required to synthesize δ-hexalactone via a PKS</td>
<td>72</td>
</tr>
<tr>
<td>3.4 Complete architecture for a putative modular megasynthese chassis</td>
<td>72</td>
</tr>
<tr>
<td>3.5 Proposed mechanism for δ-hexalactone biosynthesis from a PKS</td>
<td>73</td>
</tr>
<tr>
<td>3.6 Histogram of codon adaptation indices (CAIs) for modular megasyntheses</td>
<td>74</td>
</tr>
<tr>
<td>3.7 Codon usage normalized by position</td>
<td>75</td>
</tr>
<tr>
<td>3.8 Synthetic biology open language (SBOL) representation of complete construct</td>
<td>76</td>
</tr>
<tr>
<td>3.9 Approach to assembling modular megasyntheses from gBlock ‘parts’</td>
<td>77</td>
</tr>
<tr>
<td>3.10 Gel Electrophoresis of Gibson assemblies</td>
<td>78</td>
</tr>
<tr>
<td>3.11 Gel electrophoresis of junction PCRs</td>
<td>80</td>
</tr>
<tr>
<td>3.12 Yeast assembly of complete construct</td>
<td>81</td>
</tr>
<tr>
<td>3.13 Proposed mechanism for cross-over</td>
<td>82</td>
</tr>
<tr>
<td>3.14 Yeast assembly of synthesized modules</td>
<td>83</td>
</tr>
<tr>
<td>3.15 Genome integration of construct</td>
<td>85</td>
</tr>
<tr>
<td>3.16 BL21(dE3)STAR strain preparation</td>
<td>86</td>
</tr>
<tr>
<td>3.17 Label-free quantitative shotgun proteomics results</td>
<td>87</td>
</tr>
<tr>
<td>3.18 Growth curves for each expression condition</td>
<td>91</td>
</tr>
<tr>
<td>3.19 Solid-phase microextraction results</td>
<td>95</td>
</tr>
<tr>
<td>3.20 Gas chromatogram of liquid injection</td>
<td>96</td>
</tr>
<tr>
<td>3.21 Mass spectra of liquid injections</td>
<td>96</td>
</tr>
<tr>
<td>3.22 Extracted ion chromatograms (XICs) for LB culture conditions</td>
<td>97</td>
</tr>
<tr>
<td>3.23 Extracted ion chromatograms (XICs) for EZ Rich culture conditions</td>
<td>98</td>
</tr>
<tr>
<td>3.24 Mechanism of the mat operon</td>
<td>101</td>
</tr>
<tr>
<td>3.25 Flaviolin biosynthesis</td>
<td>102</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 - Complex Adaptive Systems

In the midst of a recession, what mix of monetary and fiscal policy will restore health to the economy? As urbanization continues, how should new cities be designed to yield similar dynamism to New York City or Paris? What mutations will yield a microbial strain with the right balance of ethanol production and tolerance? Or more relevant to organic chemistry and chemical biology: what proteins do we need to effect in order to eliminate Alzheimer’s Disease? What small molecules will have the appropriate effect on these proteins? Each of these unanswered questions shares an underlying characteristic: the context of the question is widely agreed to be a complex system. (Alper, Moxley, Nevoigt, Fink, & Stephanopoulos, 2006; Arthur, 1999; Coffey, 1998; Endy, 2005; Hardy & Selkoe, 2002; Oltvai & Barabási, 2002; Portugali, Meyer, Stolk, & Tan, 2012; Waldrop, 1993; Woodruff et al., 2013)

Engineering innovations built on scientific advances were a hallmark of the Industrial Age. Unearthing the elegant simplicity of the physical world enabled us to make massive leaps in quality of life. (Lindert & Williamson, 1983) However, scientists were restricted by their limited ability to process information to unravelling what turned out to be relatively simple phenomena. As the Industrial Age transitions into the Information Age and our ability to process information grows exponentially, scientists are beginning to investigate phenomena of increasing complexity. (Holland, 1992)
Naturally, as scientists from an array of disciplines began to recognize the ubiquity and essentiality of complex phenomena they began to organize and define a new field of study, complexity science (Waldrop, 1993). The ‘complex systems’ studied by complexity scientists share four properties (Solé, Manrubia, Luque, Delgado, & Bascompte, 1996):

1. They are made up of a large number of locally interacting elements
2. These elements have a relatively small number of simple behaviors
3. The system has irreducible macroscopic properties
4. The dynamics of these systems are robust to noise, failure, or removal of single elements

An important subclass of complex systems has a fifth property (Holland, 1992):

5. The elements (1 & 2) of the system can change and/or reorganize to yield improved macroscopic properties (3) for a given context while maintaining robustness (4)

In other words, there are ‘adaptive’ complex systems. Robustness makes complex adaptive systems pervasive and essential to the world as we know it. Ecosystems, firms, cities, markets, economies, and etc., are each complex adaptive systems.

1.1.1 - Complex Adaptive Systems, Chemical Biology, and Drug Discovery

In addition, all biological systems are complex adaptive systems, making complex adaptive systems especially relevant to chemical biology and organic chemistry as chemical biologists use small molecules produced via synthetic organic chemistry to study and manipulate biology. (Barabási & Oltvai, 2004; Ideker, Galitski, & Hood, 2001; Oltvai & Barabási, 2002) Often these
molecules are discovered during a disease-specific research program. (Schreiber, 1991) This blurs the distinction between chemical biology and drug discovery: frequently tool compounds discovered by chemical biologists are also tested for clinical efficacy as drugs.

Figure 1.1 | The yeast protein-protein interaction network. Each node represents a protein and each edge represents a protein-protein interaction. This is an excellent representation of how biological systems are dominated by many locally-interacting elements. (Barabási & Oltvai, 2004)

Historically, chemical biology and drug discovery have taken a target-based approach to discovering small molecules in which a disease is reduced to a single troublesome target. This target is developed into an assay compatible with high-throughput screening, and many small molecules (up to 100s of thousands) are screened against the target. (Drews, 2000) The results of
target-based drug discovery have been underwhelming, and the pharmaceutical industry is undergoing a crisis due to skyrocketing drug discovery and development costs attributed to a precipitous decline in productivity. (Cook et al., 2014; J. A. DiMasi, Hansen, Grabowski, & Lasagna, 1991; Joseph A. DiMasi, Grabowski, & Hansen, 2015, 2016; Joseph A. DiMasi, Hansen, & Grabowski, 2003; Sams-Dodd, 2005; Scannell, Blanckley, Boldon, & Warrington, 2012)

One strong hypothesis for the underwhelming results of target-based drug discovery is that due to the complexity of biological systems, many diseases cannot be reduced to single targets, or the ones that can be already have been. This implies that biology and disease must be understood at the systems level, and the key to avoiding the pharmaceutical industry’s existential threat may lie in effective methods for systems-level drug discovery. (Ahn, Tewari, Poon, & Phillips, 2006; Butcher, Berg, & Kunkel, 2004; Csermely, Korcsmáros, Kiss, London, & Nussinov, 2013; Goh et al., 2007; Hood, Heath, Phelps, & Lin, 2004; Hopkins, 2008; Loscalzo, Kohane, & Barabasi, 2007; Yıldırım, Goh, Cusick, Barabási, & Vidal, 2007)

1.1.2 - The Challenge of Rationally Manipulating Complex Adaptive Systems

From both a scientific and engineering perspective, rationally manipulating complex adaptive systems, as in systems-level drug discovery, presents a unique challenge. We know that the macroscopic properties of complex adaptive systems originate from many simple, locally interacting elements. In theory, we ought to be able to manipulate some set of elements and/or interactions to arrive at desired properties. This would be akin to a classic, reductionist engineering approach, dependent on scientific laws that can perfectly describe and predict behavior. Yet, this is also where complex adaptive systems are unique - thus far their macroscopic properties have proven irreducible. (Ahn et al., 2006; Gallagher, Appenzeller, Normile, & Service, 1999) In the
context of chemical biology, drug discovery, and medicine, at the most fundamental level diseases can be considered undesirable macroscopic properties of some biological complex adaptive systems. This explains why many diseases are so difficult to reduce to a single local element that can be targeted by a single, highly specific drug, let alone a set containing both local elements and their interactions. (Berger & Iyengar, 2009; Boran & Iyengar, 2010; Hopkins, 2007, 2008)

Given that the irreducibility of complex adaptive systems renders classical rational and reductionist engineering approaches ineffective, a more appropriate approach may involve introducing randomness into the engineering process. This would involve producing many different designs, testing each one, and then selecting for the best performance along the metric of interest. A further improvement would entail making the process iterative, repeatedly designing, building, and testing.

Of course, this is a wink towards evolution, the archetypal process for modifying complex adaptive systems. Evolution has been producing and modifying complex adaptive biological systems for billions of years and thus, could be an effective approach to systems level drug discovery. (Dodd et al., 2017) As a result of evolution’s essentiality in sculpting biological systems, biology is arguably the first field where we began to: a) learn how evolution sculpts complex adaptive systems and b) attempt to engineer complex adaptive systems.

1.2 - Evolution

Evolution is the process of gradual adaptation. The mechanism responsible for this process, natural selection, first proposed by Charles Darwin, is simple. (Darwin, 1859) First, biological systems are inherently unstable, and given enough time, every biological system changes. This instability exists because genomes randomly mutate - they are not maintained, replicated, and/or
passed from generation to generation with perfect fidelity. Mutations yield altered phenotypes, or observable properties. At the population scale (ie N>1), mutation yields a diversity of phenotypes. The diverse members of a population then compete to survive and/or reproduce. The best phenotypes have the highest survival and/or reproduction rates, and are then passed to the next generation, enriching as a percentage of the total population. Fitness is the value of this differential ability to survive or reproduce, measured as change in relative abundance between generations:

\[
\text{Fitness} = \text{Relative Abundance } T_i - \text{Relative Abundance } T_{i-1}
\]

If the fitness value for a given phenotype is greater than 0, the phenotype is under positive selection and is enriching. If the fitness value is less than 0, the phenotype is under negative selection and is depleting. A fitness value equal to zero means the phenotype is not under selection.

Fitness is a function of both genotype and environment. For a given genotype, as environment changes, so will fitness. For example, an *E. coli* auxotroph will have wildly different fitness values under different culture conditions. It will reproduce well in a rich media such as LB but won’t reproduce at all in a minimal medium like M9. (Davis & Mingioli, 1950) A plot of fitness as a function of environment for a fixed genotype is called a norm of reaction. (Griffiths, Miller, Suzuki, Lewontin, & Gelbart, 2000) The opposite is also true, for a given environment, each genotype has a fitness value. A useful theoretical construct to visualize this concept is the fitness landscape, in which the entire design space of genotypes and their respective fitnesses are plotted, yielding a ‘landscape’ of peaks and valleys, with peaks representing local and/or global fitness maxima. (Orr, 2005; Wright, 1932)

Of course, in reality the design space of potential genotypes is infinite, and the fitness landscape is N-dimensional. (Orr, 2005) Thus, the probability that a population contains a
genotype at the global fitness maximum is zero. Natural selection only operates on an infinitesimally small portion of the landscape during any given generation. This problem is overcome via iteration, in which natural selection repeats ad infinitum yielding slow optimization over time or, evolution.

Over billions of years, life on earth has encountered innumerable environmental threats and opportunities. Slow optimization via evolution has overcome or exploited each and every one, yielding life as we know it. In On the Origin of Species, Charles Darwin outlines numerous examples of this. (Darwin, 1859) ‘Darwin’s finches’ are particularly well known. (Lack, 1947)

![Figure 1.2 | Darwin’s finches](image)

Darwin’s drawings of four different finch species, with an emphasis on their differing beak styles. (Darwin, 1889)
In short, the ancestral finch species that populated the Galapagos Islands evolved into different species, each populating a distinct ecological niche. As it turns out, these finches have a diverse array of beak morphologies, with each beak having optimized over millions of years for a specific food source. (Lamichhaney et al., 2015) This sort of evidence was enough to inform the development of Darwin’s theory however, he did identify limitations to his work. In chapter VI he outlines some potential objections to his theory of evolution. For example, there is little evidence for ‘transitional forms’ of species. Why were there only 13 species of finch on the Galapagos Islands and not many more with intermediate beak morphologies?

1.2.1 - Uncovering the Properties of Evolution

Darwin resolves the paradoxical lack of transitional forms by impugning the quality of the fossil record:

"The geological record is extremely imperfect, and this fact will to a large extent explain why we do not find interminable varieties, connecting together all the extinct and existing forms of life by the finest graduated steps, He who rejects these views on the nature of the geological record, will rightly reject my whole theory."

(Darwin, 1859)

This issue could be explained away if there were only a complete fossil record in which each step over the course of an evolutionary trajectory was available. Paleontologists have uncovered numerous fossils since the 19th century but it is still impossible to know, a priori or posteriori, whether or not a geological record is ‘complete’. Preeminent evolutionary biologist and
paleontologist Stephen Jay Gould described how paleontologists failed to identify Darwin’s ‘transitional forms’, and how this problem haunted their field:

“Paleontologists have paid an exorbitant price for Darwin's argument. We fancy ourselves as the only true students of life’s history, yet to preserve our favored account of evolution by natural selection we view our data as so bad that we never see the very process we profess to study.” (Stephen Jay Gould, 1977)

Darwin’s and Gould’s comments both illuminated a broader challenge when it comes to transitioning evolution from scientific theory to scientific law (Gould’s thoughts on theory and fact notwithstanding (S. J. Gould, 1981)). There was significant evidence that evolution occurs and a theory for its mechanism (natural selection) but paltry evidence for said theory and its properties. Evolutionary biologists had more questions than answers. What are the dynamics of evolution? How long can fitness increase? Is it repeatable? Are there multiple evolutionary paths to the same solution? (Lenski, 2017) Along with believing that the fossil record was too imperfect to rely on, Darwin also believed that evolution was too slow to observe on relevant timescales (Darwin, 1859), thus precluding the collection or measurement of these properties. These beliefs have been disproven, as great leaps have been made in measuring the properties of evolution since Darwin, and even since Gould’s untimely death in 2002. (Buckling, Maclean, Brockhurst, & Colegrave, 2009)

As early as the 1880s biologists began to disprove Darwin’s assumption that evolution couldn’t be observed, beginning with William Dallinger culturing protozoa strains at gradually increasing temperature over the course of several years. (Dallinger, 1887) This is widely
considered to be the first example of evolution occurring on a laboratory timescale, also known as experimental evolution. Dallinger’s results suggested he had observed evolution but questions have been raised about the quality of his experiment, given the techniques available at the time. (Lenski, 2017) In the 1950s, Novick and Szilard observed selective sweeps in chemostatic bacterial cultures that were confirmed by competing isolated mutants against one another. (Novick & Szilard, 1950) Years later, biologists had access to suitable experimental tools to begin studying evolutionary dynamics in depth. (Conrad, Lewis, & Palsson, 2011; Dykhuizen & Hartl, 1980)

The quintessential example of characterizing evolutionary dynamics via experimental evolution is Richard Lenski’s Long-Term Evolution Experiment, or ‘LTEE’. (Elena & Lenski, 2003; Lenski, Rose, Simpson, & Tadler, 1991) Briefly, Lenski’s lab has grown *Escherichia coli* at 37°C in glucose limited, defined minimal media for over 50,000 generations. (“Celebrating 50,000 Generations of the Long Term Lines,” n.d.) Over the course of these 50,000 generations, they have monitored *E. coli*’s adaptation to this environment, making a large number of interesting discoveries about evolutionary dynamics along the way. (Blount, Barrick, Davidson, & Lenski, 2012; Blount, Borland, & Lenski, 2008; Gerrish & Lenski, 1998; Lenski et al., 2015; Sniegowski, Gerrish, & Lenski, 1997; Wiser, Ribeck, & Lenski, 2013)

The discoveries made in Lenski’s LTEE are a result of elegant experimental design and powerful new tools. In many ways the LTEE and experimental evolution in general solves some of the problems originally outlined by Darwin. Here is Stephen Jay Gould again, this time describing how the LTEE gets around the difficulties that haunt paleontologists:

“[In the LTEE,] samples of the common ancestral population, and of selected stages in the history of each population, were stored at -80°C, and can be revived for
competition experiments with the continually evolving populations—a situation that can only fill the paleontologist with envy, and with thoughts of beautiful and utterly undoable experiments from life’s multicellular history (neanderthals or australopithecines released in New York City; tyrannosaurs revived to compete against lions in a field of zebras, etc).” (Stephen Jay Gould, 2002)

Experimental evolution, combined with techniques in genetics allows the observation of evolution in action. (Conrad et al., 2011) Advances as simple as cryogenic storage of bacteria allowed the preservation of a perfect fossil record that can be explored in depth to elucidate the properties of evolution. The remaining limitation of experimental evolution is that for practical reasons (ie time), it is limited to the study of microorganisms and more specifically, to changes in allele frequencies within a population. This is sub-species level evolution, or ‘microevolution’. The time scale for species-level evolution, or ‘macroevolution’ is much longer, precluding its study on experimental timescales. Whether or not observations of microevolution can be extrapolated to macroevolution is still under debate. (Erwin, 2000; Reznick & Ricklefs, 2009; Simons, 2002)

Beyond accumulation of evidence for evolution itself, experimental evolution became even more impactful when combined with second-order implications of evolution. The concept of evolution shown a new light on Earth’s biodiversity by opening the door for the concept of common descent, that all life on Earth evolved from a single last universal common ancestor (LUCA). (Doolittle, 1999; Koonin, 2003; Theobald, 2010) The implications of common descent were extraordinary, as LUCA meant that all life on earth is simply an accumulation of evolutionary innovations. As scientists studied evolution in a laboratory setting, they realized that biological
innovations were also occurring on laboratory timescales. (Mortlock, 1984; G. B. Muller & Wagner, 1991; Wagner, 2011)

1.3 - Evolution as a Biological Engineering Paradigm

The observation that evolution innovates on laboratory timescales slowly developed into a new engineering paradigm, in which evolution is applied to engineer complex adaptive, biological systems. As with most significant advances, it also introduced a new set of challenges. Consider Darwin’s finches again. This time, assume one is given a finch as a template and is asked to design 13 finches with different beaks, each one appropriate for a specific food source. There is a catch however, the changes made need to be heritable - each change must be made at the genome level. The finch genome is 1.07 billion base pairs, with 16,286 protein coding genes. (Lamichhaney et al., 2015) Assuming we maintain genome size (i.e. no indels), that is a design space of $4^{1.07 \text{ billion}}$ members, effectively infinite. As such, it took millions of years for finches to accumulate the appropriate mutations that resulted in the beak morphologies seen today. (Lamichhaney et al., 2015) Although there aren’t actually any biologists using evolution to engineer birds, this is a good representation of the challenges faced when attempting to engineer using evolution.

To successfully achieve a phenotype of interest, engineers must use the same approach that natural evolution does - modify genotypes to create a diversity of phenotypes and then apply a selective pressure that forces these phenotypes to compete against one another. This means guiding evolution using mutations to generate genotypic diversity, an appropriate selective pressure, or both. The results of evolutionary engineering are a function of these two factors, as well as time. Without enough diversity or too strong a selective pressure, an experiment may take an untenably
long time. If the selective pressure is too weak, the experiment may never converge on a single solution.

1.3.1 - Adaptive Laboratory Evolution

In the very early days of engineering via evolution, artificially introducing mutations to increase genotypic diversity was all but impossible besides simply increasing the rate of random mutation. (Adelberg, Mandel, & Ching Chen, 1965; H. J. Muller, 1927) Thus, engineers were forced to use selective pressure to guide evolution towards the desired phenotype. A simple example of this is an engineer looking to produce a biofuel such as ethanol in *E. coli*. Ethanol is toxic to *E. coli* at high concentrations, so without improved ethanol tolerance *E. coli* can only produce ethanol at limited titers. To improve tolerance, an engineer can introduce a selective pressure by growing *E. coli* under conditions with high but sub lethal concentrations of ethanol. In theory, evolution adapts *E. coli* to these conditions by selecting for strains with improved ethanol tolerance. (Goodarzi et al., 2010; Woodruff et al., 2013) This technique is commonly known as Adaptive Laboratory Evolution (ALE). (Dragosits & Mattanovich, 2013; Portnoy, Bezdan, & Zengler, 2011)

In 1984, Robert Mortlock outlined what may be the earliest conception of ALE, describing a series of evolutionary experiments in which bacteria developed novel metabolic capabilities via evolution as a new form of genetic engineering. His examples are experiments in which species from the genera *Klebsiella* and *Escherichia* were grown under conditions that included reduced five carbon sugars that neither could metabolize as their sole carbon source. Given an inoculum with a large enough population (i.e. diversity) and enough time, both genera evolved the ability to metabolize the reduced sugars. (Mortlock, 1984) These experiments laid the foundation for
applying evolutionary approaches to modern engineering challenges. (Jiang et al., 2014; Rubin, 2008; Sandoval, Mills, Zhang, & Gill, 2011; Steen et al., 2010; Zhang et al., 2009)

There are significant limitations to ALE. For one, it relies on natural random mutation to create diversity. This limitation is dramatic, limiting engineering to the strain level. As engineering goals become more targeted, such as engineering a specific protein to have increased function, an ALE approach is difficult because mutations accumulate across the entire genome, rather than just in the gene of interest. In addition, if the initial selective pressure is too strong, there may not be enough initial diversity to find a solution as natural mutation is fairly rare. (H. J. Muller, 1927) Adequate diversity explains why the experiments described by Mortlock required large inoculums. For these reasons, as the complexity of the phenotype of interest increases, the less likely it becomes for the phenotype of interest to arise during an ALE experiment.

Also, as described, ALE is dependent on selection. Selection is useful because any number of genotypes can be forced to compete against one another to survive and reproduce in a given environment. In ALE experiments, survival and reproduction yield enrichment via improved growth rate. Some phenotypes however, are very difficult to tie to growth rate. In fact, some phenotypes have proven antagonistic to growth. Take the simplified biofuel example posed above but instead of tolerance, assume the goal is to improve production of ethanol in a wild type E. coli strain. Ethanol production requires resources that could otherwise have been used for growth. As a result, mutants that lose the ability to produce ethanol can funnel those resources into growth, outcompeting the mutants producing high titers of ethanol instead of rapidly growing. A simple selection-based experiment would yield strains with little to no ability to produce ethanol.

1.3.2 - Directed Evolution
In the late 1980s and early 1990s, a Cambrian explosion of tools provided engineers a route around these limitations to ALE. New molecular biology methods allowed engineers to introduce mutations in a directed way, effectively biasing the diversity in their populations toward the locus of interest. (Cadwell & Joyce, 1992; Stemmer, 1994) Around the same time, screening methods began to be introduced. In general, screening is simply a more directed form of selection in which the phenotype of interest does not have to be growth. (Daniel D Shoemaker, Deval A Lashkari, Don Morris, Mike Mittman, Ronald W Davis, 1996; Fodor et al., 1993; Smith et al., 2009) These new, more specific tools introduced a new approach to evolutionary engineering, that of directed evolution.

To this point, evolutionary approaches had been applied solely to strain-level (or genome-scale) phenotypes. As tools improved and became more directed, engineers began to focus on smaller sub-species level loci such as single genes or small regulatory nucleic acids. Theoretically, engineering should become simpler as scale shrinks, perhaps making an evolutionary approach unnecessary. Unfortunately, however, even single genes are still complex; significant non-linearities exist in the relationship between sequence and function. This phenomenon is known as epistasis. The existence of epistatic interactions means that many of the same challenges exist when engineering at the gene scale, still limiting the usefulness of a rational approach. As a result, evolutionary approaches are still considered the most appealing methods for biological engineering genes and other sub-species level loci.

One model example of directed evolution is the evolution of subtilisin E to function in high concentrations of the organic solvent dimethylformamide (DMF). (Chen & Arnold, 1993; You & Arnold, 1996) The native function of subtilisin E is as a protease in aqueous solvents. As the authors were interested in a different but narrowly defined phenotype, they required directed tools.
First, they mutagenized the exact gene of interest, subtilisin E, rather than the entire *Bacillus subtilis* genome, using a PCR-based random mutagenesis strategy. In addition, the phenotype of interest, subtilisin E activity in DMF does not have an obvious link to growth rate of *B. subtilis*, so the authors implemented a screen. Briefly, they plated the mutagenized *B. subtilis* on an organic-solvent resistant membrane and transferred this membrane to plates containing DMF. Halos suggesting expression of active subtilisin were compared to wild type controls, a few hundred colonies of interest were selected and assayed for subtilisin function in the presence of DMF. The result was a variant of subtilisin E with dramatically improved properties in DMF: a 130-fold improvement of catalytic efficiency over wild type, due to a decrease in the Michaelis constant $K_m$ (by a factor of 7) as well as an increase in the catalytic rate constant $k_{cat}$ (by a factor of 19). This outcome was attributed to a significant stabilization of the transition state in DMF. (Chen & Arnold, 1993)

In large part, directed evolution experiments have followed this strategy introduced by Chen and Arnold. The experiment begins by selecting a target functionally close to the goal phenotype. This target is then mutagenized. Mutants are screened for the target phenotype. These steps are repeated iteratively until the target has adequate functionality. (Arnold, 1998)
Figure 1.3 | Canonical approach to directed evolution. Bloom and Arnold outline a traditional directed evolution experiment. An experiment begins with a library of random (or targeted) mutations in a parent gene, this library is cloned, and then selected or screened for the phenotype of interest. For multiple cycles, winners are mutated again, and the cycle is repeated. (Bloom & Arnold, 2009)

The success of the early examples of directed evolution led to relatively widespread adoption. New mutagenesis tools and screening methods are constantly being introduced and developed. As a result, directed evolution has been applied to engineer numerous biological systems over the past 20 years and as the power of directed evolution became obvious, it quickly expanded into applications relevant to drug discovery. (Acevedo-Rocha, Agudo, & Reetz, 2014; Ceres, Trausch, & Batey, 2013; Chockalingam, Chen, Katzenellenbogen, & Zhao, 2005; Fischbach, Lai, Roche, Walsh, & Liu, 2007; Fox et al., 2007; Tuerk & Gold, 1990)

1.4 - Application of Directed Evolution to Drug Discovery
A dominant paradigm in drug discovery is inhibition of gain-of-function in the cases where gain-of-function causes disease. The autoimmune disease Crohn’s disease is a perfect example. In this case, the gain-of-function occurs in the immune system, and the overactive immune system attacks native tissue, usually in the digestive tract. Treatment for Crohn’s disease is inhibition of the immune response and currently, the highest grossing drug in the world is adalimumab or Humira, an antibody that binds and inhibits Tumor Necrosis Factor alpha (TNF-α), a cytokine, as a treatment for Crohn’s disease and other inflammatory diseases. (Hanauer et al., 2006; Weinblatt et al., 2003)

Antibodies like adalimumab are a class of adaptive humoral immune protein that have found widespread use in medicine due to their extraordinary binding affinity for almost any ligand (antigen), making them highly effective inhibitors. In a classic antibody drug discovery program, an antibody with high specificity to the given disease target must be identified. This is a protein engineering problem, and an excellent application for directed evolution. (Alberts et al., 2002a)

Much antibody discovery is still performed using animals. Briefly, an animal such as a goat, rabbit, or mouse is inoculated with the antigen, in this case the disease target. The immune system then creates a B lymphocyte (B cell) that produces and excretes an antibody which binds to the target. As these B cells are produced in the spleen, a splenectomy is performed, and the collected B cells are fused with immortal B cells. These hybridomas are screened for target antibody production via western blot. (Goding, 1980)

Unsurprisingly, the mammalian immune system uses an evolutionary approach to generate high affinity antibodies. The human body contains a pool of approximately $10^{12}$ different antibodies produced by B cells, known as the preimmune antibody repertoire. This preimmune antibody repertoire is generated by the genomic V(D)J recombination process during B cell
development. In the presence of antigen, this population is effectively subjected to a selection, as antigen binding combined with T\(_{FH}\) cell activation triggers conversion of the naive B cell to a memory B cell, with the unactivated naive B cells dying out. This represents a broad search across the fitness landscape, where fitness in this case is antigen binding. During the transition from naive B cell to memory B cell, somatic hypermutation of the antibody begins, with the high mutation rate optimizing and maximizing antibody affinity. This is a local search of the fitness landscape. Again, the B cells producing the highest affinity antibodies preferentially survive, yielding a small final population of B cells from the preimmune antibody repertoire that produce high affinity antibodies. (Alberts et al., 2002b)

For various ethical, financial, scientific, and medical reasons, using animals to generate antibodies for drug discovery is discouraged. Considerable effort and resources have been applied to development of alternative antibody directed evolution methods. The classical result of this effort is phage display. In phage display, a library of antibodies is cloned into phage, and then expressed on the phage surface. This library is washed through a column with bound antigen, and the phages producing high affinity antibodies are recovered, amplified, and the process is repeated. (Winter, Griffiths, Hawkins, & Hoogenboom, 1994)

Of course, there are also other directed evolution methods that have been applied to drug discovery. For example, Systematic evolution of ligands by exponential enrichment (SELEX) is an \textit{in vitro} method for RNA aptamer evolution. (Tuerk & Gold, 1990) Macugen is an FDA approved anti-VEGF RNA aptamer that was discovered via SELEX. (Ng et al., 2006) VEGF is implicated in cancer and other indications. Anti-VEGF antibodies have been discovered as well.

Directed evolution has been an incredibly powerful tool for engineering biomolecules and especially in drug discovery. Diversity, iteration, and strong selection, the classic evolutionary
algorithm, is highly effective at solving engineering problems in complex systems. Unfortunately however, scientists and engineers have been unable to apply directed evolution to all types of biomolecules.

Even with the massive growth in protein therapeutics, small molecules are still the dominant therapeutic paradigm but, no directed evolution approaches to small molecule discovery have been developed or implemented. This is the focus of the work presented here: developing a foundation for small molecule drug discovery via directed evolution.

1.5 - Directed Evolution of Small Molecules in Chemical Biology and Drug Discovery

As described in section 1.1.1, historically chemical biology and drug discovery have followed a target-based approach but, it is rare to successfully distill a complex disease phenotype into a single malfunctioning target. Recently, an alternative to the target-based approach has been proposed, ‘diversity-oriented synthesis’ (DOS), in which chemists attempt to synthesize large libraries of structurally diverse small molecules for phenotypic screening. (Schreiber, 2000) In theory, DOS circumvents the issues with reductionism by directly examining phenotype instead of single target binding events. This concept for DOS-based approaches to small molecule drug discovery, generation of diverse small molecule libraries followed by phenotypic screens, is strikingly similar to directed evolution.

1.5.1 - Considerations for a Small Molecule Directed Evolution Platform

Essential requirements for implementing an evolutionary approach to engineering have been outlined above. They include the following:
1. Genotypically encoded phenotypes
2. Genotypic and phenotypic link
3. A diverse population of phenotypes
4. A strong selective pressure
5. Iteration

Take the phage display example from above. The antibody is encoded by a gene, and this gene is encapsulated by the phage. Thus, the antibody itself (the phenotype) is covalently linked to the surface of the phage, maintaining the connection between phenotype and genotype so when the high affinity antibodies are collected, the correct genotypes are also collected. The antibody gene can be mutagenized by any number of techniques. Finally, the best antibodies are iteratively selected for by binding to antigen.

Historically, chemical biology and small molecule drug discovery have lacked most of these properties. Small molecules are synthesized by organic chemists, they are not genetically encoded. Thus, there is not a genotypically encoded phenotype, and there is obviously no genotype-phenotype link. As described, diversity-oriented synthesis relies on high-throughput screening, where a diverse library of small molecules is screened against the target or phenotype of interest. In some ways, this represents a diverse population of phenotypes and a selective pressure but, the many approaches for library generation have not lived up to expectations. (Feher & Schmidt, 2003) Compared to antibody libraries or RNA libraries (10^{12} members) small molecule libraries are orders of magnitude smaller (10^4-10^6). Small libraries mean a less thorough search of the fitness landscape, where in this case small molecule fitness is represented by the property of interest, for example some kind of inhibition. In addition, the winners cannot be easily amplified
for repeated rounds of screening or subjected to further mutagenesis for a local search of the fitness landscape. Developing a small molecule directed evolution platform necessitates satisfying each of these requirements.

The first step is a genetically encoded method for producing small molecules. When one begins researching small molecule drugs, it quickly becomes obvious that even after years of rational drug design and discovery, the majority of small molecule drugs are still natural products. Between 1981 and 2006, 63% of 974 New Chemical Entities were derived from natural products rather than from completely synthetic molecules. (Newman & Cragg, 2007) As natural products are structurally diverse, have high biological activity, and must be genetically encoded, they represent an ideal starting point for small molecule directed evolution.

1.6 - Modular Megasynthases, Combinatorial Biosynthesis, and Directed Evolution

Many FDA approved small molecules have come from a class of natural product secondary metabolites known as polyketides. Like most natural products, polyketides are structurally diverse and have high biological activity. (Staunton & Weissman, 2001)
Figure 1.4 | Polyketide structures. A few examples of bioactive polyketide natural products, both toxic and therapeutic. (Staunton & Weissman, 2001)

The earliest study of polyketide biosynthesis revealed that the corresponding genes are highly modular and hierarchical, producing megasynthases that perform much like an assembly line. (Donadio, Staver, McAlpine, Swanson, & Katz, 1991) In that study, three open reading frames were discovered that encoded pairs of covalently-linked synthetic modules. The order of these genes in the genome directly corresponds to the order in which the synthesis occurs. At minimum, each module contains covalently-linked ketosynthase (KS), acyltransferase (AT), and acyl-carrier protein (ACP) domains. There are also optional tailoring domains such as ketoreductase (KR),
dehydratase (DH), and enoylreductase (ER) domains that incorporate different functionality into the polyketide. In addition, there is a distinct logic to the organization of catalytic domains within each module, beginning with a KS, followed by an AT, the three optional tailoring domains, and finally, the ACP. (Fischbach & Walsh, 2006; Hertweck, 2009)

Figure 1.5 | Modular megasynthase architecture. This is the architecture of the polyketide synthase (DEBS) that produces 6-deoxyerythronolide B (6-dEB), the macrocyclic core of erythromycin. DEBS was shown to be made up six modules that condense propionyl-CoA and methylmalonyl-CoA to 6-dEB. (Donadio et al., 1991)

A reaction on a PKS begins on the loading module with an AT loading an acyl-CoA derivative onto an ACP. In the next module (the first extension module), a KS then condenses the acyl-CoA derivative on the loading ACP with the acyl-CoA derivative on the next ACP down the chain, generating a ketide. The carbonyl in this ketide then undergoes various reductions, depending on the reductive tailoring domains present in the module. For example, a KR reduces the ketide to an alcohol. A KR and DH reduces the ketide to an alcohol and then reduces the alcohol to an alkene. A KR, DH, and ER produces a fully reduced hydrocarbon. The ketide produced by a PKS is dependent on the number of modules in the enzyme and the domain structure of these modules. Finally, a thioesterase (TE) hydrolyzes the ketide from the enzyme, preferably intramolecularly
using an alcohol, generating a lactone or, with water, generating an organic acid. (Staunton & Weissman, 2001)

Nonribosomal peptide synthases (NRPSs) are another class of enzyme that has the same modularity, hierarchical architecture, and logic as PKSs. (Fischbach & Walsh, 2006; Marahiel, Stachelhaus, & Mootz, 1997) The only difference between the two classes is that instead of acyl-CoA derivatives, NRPSs use adenylated amino acids as their substrates. Using amino acids dictates that NRPSs contain different catalytic domains within modules. Instead they have an adenylation domain (A) that adenylates and loads an amino acid onto a peptide carrier protein (PCP) and a condensation domain (C) that condenses amino acids from two PCPs.

Enzymes have also been discovered that contain modules of both PKSs and NRPSs and are known as PKS-NRPS hybrids. (Du & Shen, 2001) Together, PKSs, NRPSs, and PKS-NRPS hybrids are a class of enzymes known as modular megasynthases.

The simplicity of the logic found in modular megasynthases suggested they may also be easily programmable, giving molecular biologists access to the biological activity and structural diversity of natural products. (Katz, 1997; Khosla, 1997) The idea of a library containing modular megasynthase genes with many combinations of shuffled modules and catalytic domain addition or loss of function was termed ‘combinatorial biosynthesis.’ (Rohr, 1995; Tsoi & Khosla, 1995)

Combinatorial biosynthesis of modular megasynthases represents an ideal platform for directed evolution of small molecules. Modular megasynthases are genetically encoded, the genotype and phenotype are linked by cellular encapsulation, and they are capable of producing complex small molecules.

1.6.1 - Previous Work in Combinatorial Biosynthesis
In 2005, Menzella et al. produced the seminal work in combinatorial biosynthesis, and clearly described the associated logic. Small molecule diversity can be produced by manipulating modular megasynthases in three ways: one, by adding or removing entire extension modules from the PKS, influencing the size of the small molecule (chassis length); two, by altering the reduction domains to completely reduce, partially reduce, or not reduce each acyl unit, influencing the functional groups present on the small molecule (chassis structure); and three, by altering the specificities of the acyl transferases that load each module, influencing the structure and functionality of the small molecule (chassis specificity). Their study produced only 72 of 154 molecules targeted in their combinatorial library. In order to properly explore chemical space in the course of a directed evolution experiment, a small molecule library would have to be orders of magnitude larger. A 42% efficiency would make achieving this scale exceedingly difficult.

Figure 1.6 | Structure of a PKS extension module. Presented on the left is the architecture of the Pikromycin PKS, with extension module 5 highlighted. On the right is the three-dimensional structure of this module, as collected with Cryo-EM. (Dutta et al., 2014)
In addition, Menzella et al. describe the complexity of experimental design for a system of this magnitude. Working in non-model organisms with immense (10kb-100kb) pieces of DNA forced them to transfer genes into the model organism *Escherichia coli*. They were largely restricted to standard cloning and assembly technologies that are drastically less efficient than recently developed single-step homology-based methods, such as Gibson Assembly or Circular Polymerase Extension Cloning. (Gibson et al., 2009; Kodumal et al., 2004; Quan & Tian, 2009) In addition, proteins with the size and complexity of these megasynthases are notoriously difficult to express in *E. coli*. (Pfeifer, Admiraal, Gramajo, Cane, & Khosla, 2001) The past decade has allowed for the development of a deeper understanding of the importance of everything from codon usage to translation rates in folding of PKSs. (Angov, Hillier, Kincaid, & Lyon, 2008; Angov, Legler, & Mease, 2011; Stevens, Hari, & Boddy, 2013) Perhaps most importantly, the authors had to include all their diversity up front (ie in the gene assembly steps) because genome editing, and *in vivo* mutagenesis techniques were relatively undeveloped at this point. (Sharan, Thomason, Kuznetsov, & Court, 2009) This litany of limitations increased the complexity required of their designs and assemblies, limiting their throughput and thus, their productivity.

Largely due to limited tools for designing, building, and testing these enzymes, combinatorial biosynthesis has not reached its full potential and somewhat faded as a field. The rise of synthetic biology over the past decade has stimulated the creation of new fields of study and reinvigorated fields previously thought fallow. (Andrianantoandro, Basu, Karig, & Weiss, 2006; Endy, 2005) Combinatorial biosynthesis is an example of a field that has experienced a dramatic reemergence due to synthetic biology. (Cummings, Breitling, & Takano, 2014; Kim, Moore, & Yoon, 2015; Poust, Hagen, Katz, & Keasling, 2014; Smanski et al., 2016; Yuzawa, Keasling, & Katz, 2017; Yuzawa, Kim, Katz, & Keasling, 2012) Conceptually, the goal of
synthetic biology is to design sets of interchangeable parts from natural biology to assemble into systems that function in novel and useful new ways. (Benner & Sismour, 2005) The ideal, but as of yet unachieved, implementation of large combinatorial biosynthetic libraries fits perfectly within the established framework of synthetic biology, as it begins with a minimal set of parts (ie catalytic domains or modules) which could then be hierarchically assembled into all possible modules, not simply the ones that appear in nature. These modules could then be assembled into genes with additional diversity generated by introduction of paralogous catalytic domain ‘parts’ having varied substrate specificities. (Menzella et al., 2005; Menzella, Carney, & Santi, 2007) Unfortunately, a highly efficient, scalable approach to modular megasynthase part design and assembly for combinatorial biosynthesis does not yet exist.

1.7 - This Work: Designing Modular Megasynthases for Combinatorial Biosynthesis

We set out to design modular megasynthases via a synthetic biology-style approach: using computation to abstract away the complexity inherent in modular megasynthases, designing a set of context-independent parts that enable facile, scalable, and combinatorial assembly of modular megasynthases. To achieve this goal we first built a computational design pipeline for context-independent linker sequences (Chapter 2) that, when combined with the appropriate catalytic domains using homology-based cloning techniques, are scarlessly assembled into modular megasynthases, producing novel molecules in *Escherichia coli* (Chapter 3). Initially, I tested this approach via *de novo* assembly of modules comprising a Polyketide Synthase (PKS) that produces \textsuperscript{TM}-hexalactone, a triketide lactone food and fragrance ingredient not naturally produced by bacteria (Chapter 3). This approach is then applied to other classes of modular megasynthases: Non-ribosomal Peptide Synthases (NRPSs) and NRPS-PKS hybrids, to demonstrate the extensibility of
the design approach (Chapter 4). I also show that the designs allow for the application of massively multiplex genome engineering tools to modular megasynthese engineering, providing a scalable, high efficiency approach to combinatorial biosynthesis (Chapter 4).

1.8 References


Chapter 2

Computational Part Mining for Modular Megasynthase Engineering

2.1 - Introduction

As described in Chapter 1, two properties define synthetic biology projects: a set of parts that behave predictably, and rules for how these parts can or cannot be used in combination. (Endy, 2005) Applying synthetic biology to modular megasynthase engineering for combinatorial biosynthesis requires that these properties are met.

2.1.1 - Design Rules

The rules for designing modular megasynthases have been well established and appropriately implemented in previous attempts at combinatorial biosynthesis. Each class of modular megasynthase has a distinct set of catalytic domains that are required to form a minimal module. At minimum, polyketide synthases (PKSs) require a ketosynthase (KS), an acyltransferase (AT), and an acyl carrier protein (ACP). The canonical organization follows this order, a minimal module is [KS]-[AT]-[ACP]. Nonribosomal peptide synthases (NRPSs) require a condensation domain (C), an adenylation domain (A), and a peptide carrier protein (PCP). The canonical organization of NRPSs is [C]-[A]-[PCP]. There are also additional tailoring domains that can be included in modules, for example, ketoreductases (KR), dehydratases (DH), enoylreductases (ER), o-methyltransferases (oMTs), n-methyltransferases (nMTs), epimerization domains (E), cyclization domains (Cy), etc. In both PKSs and NRPSs, connected modules perform iterative
condensations of either acyl-CoA derivatives (in PKSs) or aminoacyl-AMP derivatives (in NRPSs). (Fischbach & Walsh, 2006; Staunton & Weissman, 2001) The unsolved problem that has likely hindered combinatorial biosynthesis is that no coherent set of ‘parts’ that behave in a reproducible, constructive way exists.

2.1.2 - Parts

The catalytic domains of modular megasynthases have been relatively well studied and understood, as outlined in Chapter 1, making them the ideal starting point for a set of modular megasynathase parts. However, catalytic domain parts alone were not enough to enable successful combinatorial biosynthesis. Clearly, appropriately defining parts required a deeper understanding of modular megasynathase mechanisms.

The studies unraveling catalytic domains came quickly after the first modular megasynathase discovery, including three dimensional structures produced for many individual catalytic domains but never for a complete module. As a result, for over 20 years the biggest open questions in the biochemistry and engineering of modular megasynthases have been understanding: one, the structure of complete modules and two, the dynamics between catalytic domains within a functional module.

In 2014, the answers to these questions began to be revealed, with a structure of a full-length module from *Streptomyces venezuelae*. (Dutta et al., 2014) Using this structure, the authors were able to study the dynamics of the enzyme over the course of a reaction. They found extraordinary dynamics, as the enzyme dimerizes to form a kind of ‘reaction chamber’ in which the ACP moves from catalytic domain to catalytic domain. In addition, they identified dynamics that suggest that each step in the catalytic operation of the module initiates the following step.
They identified that their observations, “ha[ve] profound implications for the design of PKS modules.” (Whicher et al., 2014) In 2016, another step forward was made, although this time with an iterative-PKS rather than an assembly line-PKS. This study by Herbst, et al. achieved high enough resolution to completely dissect the mechanism of a fully reducing iterative-PKS module. (Herbst, Jakob, Zähringer, & Maier, 2016) Their results provided confirmation of the dynamics proposed by Dutta et al. in 2014, attributing much of the dynamics to a linker-based architecture:

“The linker-based architecture supports modularity of the modifying region by requiring only the adaptation of variable linker regions for evolutionary domain shuffling. It thus rationalizes an important aspect of the notable success of the PKS architecture in the generation of chemical diversity. Our results highlight the relevance of matching linker–domain, rather than domain–domain, interactions in PKS engineering. They contribute to the fundamental understanding of PKS architecture, as well as to the functional dissection and re-engineering of related synthases including relevant drug targets and important producers of bioactive compounds.” (Herbst et al., 2016)

This notion aligns well with earlier theoretical studies considering the origin of modular megasynthases. These studies proposed that recombination of functional catalytic domains created inter-domain amino acid sequences which eventually evolved into linkers conferring the requisite structure for constructive interactions between catalytic domains and modules. (Gokhale & Khosla, 2000; Gokhale, Tsuji, Cane, & Khosla, 1999; Jenke-Kodama, Sandmann, Müller, & Dittmann, 2005)
These studies demonstrate how complex modular megasynthases are, and thus, how important design is. As outlined by Menzella, it is necessary to connect catalytic domains in the appropriate orders. (Menzella et al., 2005) It is also clear however, that the manner in which these catalytic domains are connected is equally essential. Even with functional catalytic domains, ineffective connections that do not allow appropriate spatial organization or dynamics likely impede constructive interactions between catalytic domains, yielding a non-functional module. Catalytic domains and linkers both appear to be essential for the function of modular megasynthases and thus define the parts needed for a synthetic biology approach to modular megasynthetic engineering. Compared to catalytic domains, linkers have been relatively neglected in studies of modular megasynthases and as such, the focus of this work is on linker design.

2.1.3 - Considerations for Linker Design

Synthetic biology requires that these linkers be context-independent, or promiscuous. In other words, ideal linkers will function appropriately regardless of the design of the rest of the modular megasynthetic. Identifying promiscuous linkers represents a significant challenge. Numerous studies of protein evolvability show that many proteins begin with promiscuous function(s), and as they evolve there is initially a weak trade-off between evolution of a new function and loss of the native function. The promiscuous function can begin to evolve without significant loss of native function. Over time however, as the new function continues to improve and optimize, the structural plasticity that allowed for promiscuous function is lost, and the native function of the protein is lost with it. (Khersonsky & Tawfik, 2010) Speculating about the evolution of modular megasynthases, it is possible that when functional modular megasynthases first appeared, the linkers were fairly promiscuous and context-independent but as certain
megasynthases evolved to produce fitness conferring secondary metabolites, the linkers evolved, trading off promiscuity for improved properties like kinetics or stability of the adaptive megasynthase. Developing a design approach for context-independent linkers demands a deeper understanding of the minimal features required for their function.

There are numerous potential biochemical approaches that could be used to unwind the function of these regions, such as an alanine scan or some other form of deep scanning mutagenesis. (Cunningham & Wells, 1989; Fowler & Fields, 2014) Unfortunately, these methods are highly context-dependent. In other words, the information revealed by this kind of approach is specific to the experimental conditions. An alanine scan of a linker sequence mostly informs about that specific linker, not the entire class of linker. In theory, some of the information likely would hold across the entire class but this isn’t knowable until the experiment has been performed across many different contexts. Thus, a targeted biochemical approach to identifying essential properties of linkers is unlikely to reveal designs with the ideal promiscuity or context-independence.

There are, however, newer computational paradigms in biochemistry that leverage large datasets to understand patterns across entire classes of proteins. Evolutionary biochemistry and phylogenetics are good examples. (Harms & Thornton, 2013) As it turns out, evolutionary approaches are useful for elucidating patterns in structure and function across entire protein classes as,

“[p]rotein evolution has been a massive experiment, conducted in parallel over billions of years, in the diversification and optimization of structure and function. The data from this experiment persist in the patterns of conservation and variation in present day sequences.” (Harms & Thornton, 2013)
Besides this use in pure biochemistry, evolutionary and phylogenetic methods have also previously been shown to have efficacy in protein design. (Chen et al., 2010; Cole & Gaucher, 2011a, 2011b)

Thus, I sought to use an evolutionary approach to design linker sequences with the appropriate properties, promiscuity and context-independence.

**2.1.4 - This Work**

All evolutionary biochemistry or phylogenetic analyses begin from a database of sequences. As no modular megasynthase linker sequence database exists, the work reported here begins with construction of a genome-mining tool to identify these sequences in bacterial genomes. Application of this tool yielded the necessary database of sequences. I then performed various phylogenetic analyses to study these databases, eventually identifying linker sequences that could be used as designs in a modular megasynthase engineering program (Figure 2.1).

![Figure 2.1 | Computational workflow for identifying context-independent linkers. Initially, complete bacterial genomes are pulled from NCBI. Putative modular megasynthases are annotated using antiSMASH, and linker sequences are identified and classified. Phylogenetic analysis is performed on each linker database, resulting in the selection of context-independent linkers.](image-url)
sequences are inferred from annotations. Linkers are classified by catalytic domains linked and then analyzed to identify a single design.

2.2 - Results

2.2.1 - Genome Mining for Linker Sequence Database Acquisition

One way to build databases of linker sequences was via literature search, identifying as many characterized modular megasynthases as possible, and collecting the linker sequences present. This of course would have been immensely time consuming, and likely would not have led to particularly large databases.

Instead, we sought a more nuanced, efficient, and scalable approach. An automated computational tool was both ideal and feasible given that machine learning has proven extraordinarily effective at pattern recognition, and dramatic improvements in machine learning methods have enabled their application to all kinds of challenges in biology, from genome annotation to text-mining the scientific literature. (Moriya, Itoh, Okuda, Yoshizawa, & Kanehisa, 2007; Seemann, 2014; Tanabe et al., 1999; Tarca, Carey, Chen, Romero, & Drăghici, 2007)

The antibiotics and Secondary Metabolite Analysis Shell (antiSMASH) represented the perfect tool for these purposes. antiSMASH is a bioinformatics pipeline for identification and annotation of biosynthetic gene clusters. Briefly, using profile Hidden Markov Models (pHMMs), antiSMASH identifies biosynthetic gene clusters in unannotated genome sequences, classifying them, and then, in the instance of PKSs and NRPSs, also identifying and annotating catalytic domains. (Blin et al., 2013; Blin, Wolf, et al., 2017; Medema et al., 2011; Weber et al., 2015)

antiSMASH was implemented to identify each putative Polyketide Synthase (PKS), Non-Ribosomal Peptide Synthetase (NRPS), and PKS-NRPS hybrid across every complete bacterial
genome in the NCBI GenBank genome database. Only complete genomes were chosen for annotation so as to maximize the quality of the data used to identify linkers. antiSMASH is capable of identifying numerous classes of biosynthetic gene clusters but it was directed to only search for Type I PKSs, NRPSs, and PKS-NRPS hybrids. When this genome mining was performed, there were just over 1,000 genomes in the NCBI Bacterial Genome database, with approximately 5,000 chromosomes. The antiSMASH output for each genome was a GenBank file with its original annotations, as well as those produced by antiSMASH. From each GenBank file, the genes of interest were identified, and the catalytic domain annotations were tested to confirm their architecture followed the canonical architecture rules. After removal of duplicates, this genome mining yielded 2,973 PKS, NRPS, or PKS-NRPS hybrid genes, made up of over 25,000 catalytic domains. The workflow for implementing antiSMASH for linker mining is presented in Figure 2.2.
Figure 2.2 | **Computational workflow for linker mining and design and complete gene design.** Genbank files of complete bacterial genomes in the NCBI database are annotated with antiSMASH to identify modular megasynthases. Linkers are inferred from these annotations. Each class of linker is clustered at 50% similarity using UCLUST. MEME and MAST are used to identify conserved motifs and score each sequence in the largest cluster. The highest scoring sequence is used as the design. This process is repeated for all required linkers. Catalytic domains and linkers are combined to complete the amino acid sequence, which is then codon harmonized.

With this database in hand, we then built a computational tool to infer the amino acid linker sequences from the catalytic domain annotations. This tool began with each modular megasynthase gene from the database described above. It then identifies the annotated catalytic domains. Using the boundaries of each catalytic domain, it is able to infer the amino acid sequence of the linker. These linkers were classified by catalytic domains linked: for example, Ketosynthase-
Acyltransferase (KS-AT) or Acyl Carrier Protein-Ketosynthase (ACP-KS). This tool yielded a database of potential designs for each class of linker.

2.2.2 - Biological Diversity of Linker Databases

We hypothesized that linker sequences conserved across a broad range of bacterial species would be more likely to function context-independently. Thus, we performed some preliminary biological diversity analyses of these linker databases to search for evidence of this. As expected, these analyses identified conserved patterns in each database. Using KS-AT linkers as an example, the KS-AT linker database contains 2,614 individuals (Figure 2.3).

Figure 2.3 | Phylogenetic tree of Ketosynthase-Acyltransferase linkers.
A histogram of the lengths of the KS-AT linkers is clearly unimodal, with a mean length of 106, suggesting some kind of conserved pattern. (Figure 2.4)

![Histogram of Ketosynthase-Acyltransferase (KS-AT) linker lengths.](image)

**Figure 2.4 | Histogram of Ketosynthase-Acyltransferase (KS-AT) linker lengths.**

In addition, we calculated Simpson’s measure of evenness ($E_s$) to be 0.283, indicating that some of the linkers in the KS-AT database are of the same sequence. Further analysis found only 1,876 of the 2,614 linkers were unique. To determine the frequency of these overrepresented linkers, we generated a rank-abundance, or Whittaker, plot. (Figure 2.5)
Figure 2.5 | Whitaker plot of KS-AT linker sequence enrichments. Every unique linker sequence is treated as a species. The dashed line represents abundance two standard deviations above the mean abundance. All species marked in green have abundance over two standard deviations above the mean.

If each species were equally represented, this plot would be flat, with each species representing $0.038\%$ of the database. The mean abundance was $0.065\%$ with standard deviation of $0.11\%$. 23 species had abundances over 2 standard deviations above the mean, with the most abundant at $2.14\%$, or a 46X enrichment. Taken together, this data suggests that each linker database was biased toward a small number of Operational Taxonomic Units (OTUs). As the linker databases are agnostic to bacterial species and enzyme context, the enriched OTU must be present in varied contexts (ie different bacterial species, enzymes, and modules), as desired.
2.2.2 - Selecting Linker Designs from Each Database

The final step in the process of computational linker design requires processing the information in each database in order to yield a linker design. To do this we first used clustering analysis as implemented in UCLUST. (Edgar, 2010) Clustering allowed us to not only identify the biased OTU but also all other linkers within the database with sequence homology to this OTU. Continuing with the KS-AT example, the complete 2,614 sequence database was divided into 961 clusters with a mean relative abundance of 0.1%. 17 clusters were over 2 standard deviations above the mean relative abundance. The largest cluster contained 117 sequences, or 5.45% of the entire database, an almost 55X enrichment (Figure 2.6).
Using the largest cluster from the database I then had to select a single sequence as the final design. This required some kind of scoring methodology. Multiple Expectation Maximization for Motif Elicitation (MEME) analysis uses a form of machine learning, expectation maximization, to identify the conserved sequence motifs within a set of sequences. (Timothy L. Bailey & Elkan, 1995; Timothy L. Bailey, Williams, Misleh, & Li, 2006) Given a set of motifs (such as those generated by MEME) and a target sequence, the Motif Alignment and Search Tool (MAST) calculates the statistical significance of motif matches to the target sequence. (Timothy L. Bailey et al., 2009; T. L. Bailey & Gribskov, 1998) This represents an ideal way of profiling the conserved
sequences within the cluster, and then assigning each sequence a score (i.e. statistical significance). Thus, the cluster was analyzed for conserved motifs using MEME, and MAST was then used to score each member of the subset. MEME identified five motifs of length 46, 43, 12, 31 and 9, respectively. MAST scored each sequence, with the highest scoring, QEVRPAPGQGLSPAVSTLVVAGKTMMQVRVSATAGMLADWMEGPGADVADVHLAHNL HHRSPQPKFGTTVVARDRTQAIAGLRLAAGQHAPGVNPAGSPGPGTVF, having an E-value of 1.110-114. This amino acid sequence was used as the final KS-AT linker design. This approach is represented in Figure 2.7 and was applied to identify designs for all other linker classes.

![Phylogenetic analysis for identifying a single design](image)

**Figure 2.7** | **Phylogenetic analysis for identifying a single design.** A) Drilling down into the KS-AT linker phylogenetic tree to the design sequence. Examining a complete cluster does not provide high enough resolution to see the individual branch that represents our design, so we drilled down into the clade representing the largest KS-AT cluster using MEME and MAST. B) Example MAST diagrams at each level as we drilled down to the final design. C) KS-AT linker peptide design.

### 2.2.3 - Preliminary Validation of Designs

53
Ideally, this approach would identify linkers that are functional *prima facie*. Unfortunately, this isn’t the case and, experimental characterization is resource intensive, so we sought a simple, efficient method to roughly validate our designs. To do so, we used SWISS-MODEL, a homology-based structural modelling tool. (Arnold, Bordoli, Kopp, & Schwede, 2006)

Briefly, SWISS-MODEL compares a peptide query to a database of experimentally determined 3D protein structures. Using structures with high homology to the query, it predicts the 3D structure and the quality of the prediction is scored. We hypothesized that some linkers may have been crystallized during catalytic domain studies and that homology modelling of the linkers would tell us if our linkers had structural similarities to the crystallized linkers in their WT contexts.

As an example, the highest scoring homology model of the KS-AT linker is shown in Figure 2.6. The KS-AT linker model sequence had 35.8% homology to a KS-AT linker that was crystallized as part of a crystal structure of the Acyltransferase from *Mycobacterium tuberculosis* Pks13. (Bergeret et al., 2012) SWISS-MODEL predicted that the design would have largely the same folding as the crystallized linker however, there is one potentially significant difference: an α-helix in the crystal structure is not present in our design. This suggests that our design has slightly less secondary structure, which may confer the desired context-independence. This same analysis was performed on each linker design over 30 amino acids in length, as SWISS-MODEL only works on sequences greater than 30 amino acids long. The results for each of these linker designs are presented in Figure 2.8.
The structure of the KS-AT linker design (left) was predicted using the SWISS-MODEL web tool for homology modelling. The wildtype KS-AT linker (right) was crystallized as part of a study of the Aeryltransferase from *Mycobacterium tuberculosis* Pks13. The design is predicted to have largely the same folding as the crystallized linker however, there is one potentially significant difference: an α-helix in the crystal structure is not present in our design.

2.3 - Discussion

We have demonstrated a computational design tool that yields catalytic domain linkers for modular megasynthase engineering. This tool uses genome mining to generate databases of putative modular megasynthase linker sequences. We performed this genome mining on all bacterial genomes available at the time. As mentioned in footnote 1, since we built these databases, the authors of antiSMASH have also used their tool on all available genomes. Given that their databases are a few years more recent than ours, the results are similar. Our work was performed on just over 1,000 genomes and yielded over 2,100 PKS, NRPS, or PKS-NRPS hybrid genes. Their
work was based on 3,907 genomes and yielded 9,000 PKS, NRPS, or PKS-NRPS clusters. Their database not include linker annotations.

As suggested in section 2.2.1, there are tradeoffs between using computational tools/models versus a literature search to build linker databases. Most of the linkers identified computationally are only putative, and it is uncertain whether or not they are functional. On the other hand, a literature search would identify a much smaller set of linkers. Given enough time, the number of experimentally characterized linkers will converge on the number identified computationally, improving the quality of the database.

An additional question that exists across any database generation method is how the border between catalytic domains and linkers is defined. In the case of antiSMASH, the border is completely defined by the pHMMs. In a literature search, these borders will have varied definition on an experimenter to experimenter basis. This boundary will necessarily be arbitrary and removing the experimenter to experimenter variation will require setting some kind of standard.

Each linker database is biased toward a small number of OTUs, and we used clustering analysis to identify these sequences as well as their sequence homologs. These clusters were analyzed, eventually identifying a single, ideal linker design. We chose to identify designs from these more common sequences because we speculate that they may have the desired context-independence for two reasons. One, they appear to have enough structure to perform the necessary function but two, maintain enough structural flexibility that it has the context-independence we are looking for. This result is consistent with hypotheses about protein plasticity, robustness, evolvability, and promiscuity. Namely, that as structural plasticity decreases, so does promiscuity and evolvability. (Khersonsky & Tawfik, 2010)
In machine learning terms, clustering classifies objects based on a shared feature or property. Our application of clustering classifies based on amino acid sequence similarity between the linker peptides. Peptides however, have numerous other features that in theory could have been implemented in machine learning classification such as three-dimensional structure, chemical properties, etc. Using a different feature to classify linkers or perhaps, multiple features could be another effective approach to linker design.

Our ultimate goal was to design functional linkers, so we designed our approach to identify linker designs that currently exist in nature however, there is also an argument to be made that currently existing linkers are likely too highly-evolved to maintain function regardless of context. Opposite of the discussion above, modern linkers may have actually evolutionarily optimized for their given context, giving up promiscuity. In this case, designing context-independent linkers would require unwinding evolution back to a state where the linkers are promiscuous. Methods exist to do this: namely, ancestral reconstruction. (Harms & Thornton, 2010) Ancestral reconstruction could have been applied to linker design but, would lead to designs that may or may not actually be functional.

The major gaps remaining in this analysis are: a) whether our linker designs function when used as parts in a modular megasynthesase engineering project and b) whether they are truly context-independent, or functional in any modular megasynthesase engineering project.

2.4 - Materials and Methods

2.4.1 - Genome Mining

We implemented antiSMASH 2.0 to annotate bacterial genomes. (Blin et al., 2013; Medema et al., 2011) The GenBank file for each complete bacterial genome was downloaded using
NCBI’s FTP server. Each GenBank file was annotated by antiSMASH 2.0 and saved. The resulting annotated genbank file is parsed to create a database of modular megasynthase CDSs. This database is stored as a text file, with each gene as a separate entry. Relevant features stored include source genome, forward or reverse strand, genomic position, nucleotide sequence, and amino acid sequence.

Another database is created containing every catalytic domain. The catalytic domains were stored with relevant info such as source genome, CDS, genomic position, type, etc. Each line in this file is a distinct catalytic domain.

Next, the CDS database and Catalytic Domain database are filtered to remove any potential duplicates. Duplicates were identified as performing every pairwise alignment with, duplicates defined as any sequences with over 99% sequence similarity. When a duplicate was identified, one of the two duplicates was removed. The remaining sequences were converted to a list of objects. There were a few classes of objects: CDSs, Modules, Catalytic Domains, and Linkers. CDS objects are made up of module objects. Module objects are made up of Catalytic Domain and Linker Domain objects. Linker Domain objects were defined as the unannotated regions between catalytic domains within modules. Methods were written to retrieve relevant features for each object (AA sequence, etc).

### 2.4.2 - Linker Database Generation

Using the objects and methods described above, we retrieved all the linkers from Type I PKS modules, NRPS modules, and PKS-NRPS hybrid modules that fit the canonical design rules, including module-module linkers (i.e. ACP-KS) and thioesterase linkers (i.e. ACP-TE). These linkers were binned by type in FASTA files.
2.4.3 - Biological Diversity Analyses

Phylogenetic trees were generated by clustering the FASTA files described above at 97% similarity using UCLUST. (Edgar, 2010) The centroids from each cluster were aligned using MAFFT on the CIPRES server. (Katoh, Misawa, Kuma, & Miyata, 2002; Miller, Pfeiffer, & Schwartz, 2010) RAxML was then used to generate a phylogenetic tree from the multiple sequence alignment MAFFT output. (Stamatakis, 2006) Figures were generated from the RAxML output using FigTree. (Rambaut, 2012)

Species richness was calculated by counting the number of unique amino acid sequences in each database. Simpson’s measure of evenness ($E_D$) was calculated as:

$$E_D = \frac{1}{S} \times \frac{1}{\sum_{i=1}^{S} p_i^2}$$

Where $S$ is species richness and $p_i$ is the proportion of species $i$ relative to the total number of species.

Rank abundance, or Whittaker, plots were generated by ranking each species by $p$.

2.4.4 - Identifying Linker Designs

Linker sequence database FASTA files were clustered using UCLUST at lower stringency (50%) to divide the databases into clusters of all unique amino acid sequences as well their sequence homologs. These clusters were converted to rank abundance plots by ranking each cluster by $p$, where $p_i$ is the number of species in a given cluster relative to the total number of species in the database. The most abundant cluster was defined as the cluster with the highest $p$.

The most abundant cluster was then subjected to MEME analysis. MEME was run in normal mode assuming zero or one motif occurrence per sequence, up to five different motifs of
width 5 to 100 amino acids. All other settings were left as default. MAST was performed on the entire cluster using the identified motifs and default settings. The sequence with the lowest E-value was subjected to further analysis.

2.4.5 - Homology Modelling

The three-dimensional structure of the winning sequence was modelled using the SWISS-MODEL web tool available at: https://swissmodel.expasy.org/. The sequence was aligned to the sequences in the default reference database. Homology models were generated from the highest scoring alignments.

2.5 - References


Chapter 3

Bottom-up, *de novo* design, build, and test of a novel modular megasynthese in *E. coli*

3.1 - Introduction

As outlined in previous chapters, a truly synthetic biology-style approach to modular megasynthese design requires a complete set of parts. The purpose of designing linkers (Chapter 2) was to complete the theoretical set of parts necessary to assemble complete functional modular megasynthese chassis. After design, the linkers must be tested for function. There are two formats for testing linker function: one, testing each linker independently in a validated system such as a functional wild type modular megasynthese or two, using a set of catalytic domain parts and designed linker domain parts to actually build and test a modular megasynthese. Our stated goal for this project was to use our parts to build novel modular megasyntheses. Thus, we chose the second approach to test our linker designs: designing a complete, novel modular megasynthese. The standard approach to prototyping that was used for these novel modular megasyntheses was a design, build, and test (DBT) cycle. (Ryan T. Gill, Halweg-Edwards, Clauset, & Way, 2016)

3.1.1 - Design

3.1.1.1 - Target Modular Megasynthese Selection and Design
There are two ways to go about designing a proof of concept modular megasynthase, what we call a forward approach and a reverse approach. The forward approach involves designing a target megasynthase first, and then developing a strategy for identifying the predicted metabolite. A reverse, or retrosynthetic, approach begins with a target molecule and then requires performing a retrosynthetic analysis in order to identify the requisite modular megasynthase that would produce the target.

For a proof of concept, such as demonstrating the efficacy of our part design, the reverse approach is more appropriate. Once a target molecule has been selected, retrosynthetic analysis, in the style of analyses used for total syntheses in organic chemistry, will yield the specific substrates and transformations and thus modules, catalytic domains, and linkers needed to build the modular megasynthase.

This way, the proof of concept begins with a target molecule that has the requisite properties. First, and most obvious, it can be produced by a polyketide synthase. Second, the substrates required (e.g. acyl-CoAs or amino-acyl AMPs) are present in the bacteria of interest (in our case, *E. coli*). Third, it is a novel bacterial secondary metabolite. Fourth, it has a readily accessible authentic standard, to simplify the detection process.

The necessary parts identified by the retrosynthetic analysis can be assembled to yield the amino acid sequence for the target modular megasynthase. The amino acid sequence can then be converted to a nucleotide sequence using codon optimization or harmonization.

3.1.1.2 - Ancillary Construct Design Considerations
Tuning codon usage via codon optimization or harmonization is a perfect example of an ancillary yet essential design consideration. It doesn’t directly impact the structure of the protein itself but can affect function. The canonical approach usually used in protein expression is codon optimization, in other words designing genes with only the most frequently occurring codons for each amino acid. Expression of large proteins however, such as modular megasynthases, is notoriously difficult in *E. coli*. Codon usage has been implicated in protein folding, as rare codons can slow translation rate via ribosomal pausing, giving proteins time to fold appropriately. (Angov, Hillier, Kincaid, & Lyon, 2008; Angov, Legler, & Mease, 2011; Kim et al., 2015; Novoa & Ribas de Pouplana, 2012) Thus, using codon optimization is not inherently the most prudent approach. Slower translation may lead to less overall protein but, a higher fraction folded and functional.

There are other ancillary parts that also impact function; for example, promoters and ribosomal binding sites (RBSs) that control expression levels. An entire spectrum of these exist, allowing fine tuning of expression. (De Mey, Maertens, Lequeux, Soetaert, & Vandamme, 2007; Salis, Mirsky, & Voigt, 2009) In the case of large proteins, expression can cause a dramatic metabolic burden and thus a severe growth defect or even toxicity. (Bonomo & Gill, 2005; R. T. Gill, Valdes, & Bentley, 2000) Total biomass is a factor in metabolite yields thus, selecting the right promoter and RBS pair to balance high levels of expression without harming metabolite yields is essential. The same can be said of other parts such as insulators, terminators, etc.

3.1.1.3 - Strain Design

Finally, the strain in which the construct will be cloned and expressed needs to meet specifications. First, the carrier protein catalytic domains in modular megasynthases must be
phosphopantetheinylated. This is performed by phosphopantetheine transferases (PPTases). For example, functional expression of heterologous modular megasynthases in *E. coli* requires an additional heterologous phosphopantetheine transferase, usually *sfp* from *B. subtilis*. (B. A. Pfeifer, Admiraal, Gramajo, Cane, & Khosla, 2001) Other bacteria do not have this requirement. Additionally, the target strain must produce the substrates required by the designed modular megasynthase. (Lombó et al., 2001)

### 3.1.2 - Build

With a nucleotide sequence in hand, numerous approaches are available to build the target strain. Both DNA synthesis and cloning technologies have rapidly improved in the past ten years. (Kosuri et al., 2010; Kosuri & Church, 2014) DNA of almost any length can be ordered and there are many scarless cloning methods. (Gibson, 2009; Gibson et al., 2009; Kok et al., 2014; Quan & Tian, 2009) Thus, the balance to strike is between cost and time. At long lengths, DNA synthesis can be prohibitively expensive but starting from trivially inexpensive 100bp oligos would make for a long and arduous cloning process. For a proof of concept, the specific build approach is not of the utmost importance but, if the technology is to scale, then the build approach becomes extremely important.

### 3.1.3 - Test

The necessary tests required to confirm function are fairly standard. After building the construct and cloning it into the target strain, the strain should be sequence verified. Appropriate
expression should be confirmed via proteomics. The final step in the test process, metabolomics to identify the target metabolite, is the most difficult and bears further introduction, as it has two major steps, metabolite extraction and metabolite detection. (van der Werf, Overkamp, Muilwijk, Coulier, & Hankemeier, 2007) The biggest challenge in identifying the target metabolite is that its unknown level of production impacts the extraction and detection methods selected.

3.1.3.1 - Metabolite extraction

Most metabolomics protocols begin with an extraction, in order to partially purify the metabolite of interest and simplify the matrix from which the metabolite must be detected. Production levels influence the extraction approach selected. At high levels, the metabolite may be excreted into the fermentation media, simplifying the extraction process due to the simplicity of the media matrix compared to the complex milieu of the cell itself. In addition, if the metabolite remains in the cell interior, the cells must be lysed prior to extraction, adding a layer of complexity. Rarely do extraction protocols have quantitative yields, and high production levels can also eliminate some of the concern over loss in the extraction process. If the molecule is produced at a very low level, loss during the extraction process becomes a significant problem, as it can drop concentrations below limits-of-detection, yielding false negative results. This can be overcome but only using an extraction method that includes a concentration step.

An appropriate extraction method also requires a good understanding of the target metabolite’s chemical properties. As an example, liquid-liquid or solvent extraction is a commonly used technique in organic chemistry in which two immiscible solvents (usually a non-polar organic solvent and a polar aqueous) are mixed, allowed to separate, and the layer containing the target
molecule is collected. Of course, this requires knowledge of the molecule’s solubility in various solvents in order to select the correct solvent pairing. The same challenge applies for other extraction methods such as liquid chromatography or solid phase extraction.

3.1.3.2 - Metabolite Detection

Once the extraction process is complete, the metabolite must be detected. Again, there are a variety of approaches to this, and a method must be selected that fits with the extraction performed. Of course, again, the production level of the molecule also plays a role in detection, as various techniques will have different limits of detection. The most common detection methods for metabolomics are mass spectrometry based with different inline chromatography (gas or liquid). (Dettmer, Aronov, & Hammock, 2007; Wishart, 2008)

3.1.4 - This Work

As our proof of concept, we used the Design, Build, Test cycle as outlined to construct a PKS capable of producing a target triketide lactone food and fragrance ingredient not known to be produced by bacteria, delta-hexalactone. A retrosynthetic analysis identified catalytic domains, modules, and substrates required for delta-hexalactone biosynthesis via a Type I PKS. These results were used to design a PKS that was then assembled, integrated into the E. coli genome, and then tested for function (Figure 3.1).
Figure 3.1 | Molecular biology workflow to produce a target strain. Catalytic domains and linkers are combined to generate a complete amino acid sequence that is codon harmonized, synthesized in pieces, cloned, and integrated into the *E. coli* genome.

3.2 - Results

3.2.1 - Target molecule selection and retrosynthetic analysis.

We selected delta-hexalactone as the target molecule as it is an industrially relevant coconut-scented food and fragrance ingredient with an estimated ~21 kg consumed annually worldwide. (Burdock, 2016) It is also a lactone, suggesting that it could be produced by a PKS and thus, via fermentation rather than chemical synthesis.

Retrosynthetic analysis using canonical rules for PKSs indicated that delta-hexalactone could be biologically synthesized using a trimodular PKS (Figure 3.1). The initial transformation is an esterification performed by a thioesterase domain (TE). The resulting synthon undergoes a
functional group interconversion to generate the canonical beta-ketoacyl functional group. This ketide is fragmented, generating a synthon equivalent to malonyl-CoA and a diketide. This diketide undergoes a functional group interconversion to generate another beta-ketoacyl functional group and then is fragmented into two synthons, one equivalent to malonyl-CoA and one to acetyl-CoA.

Figure 3.2 | Retrosynthetic analysis of 6-hexalactone. The initial transformation is an esterification performed by a thioesterase domain (TE). The resulting synthon undergoes a functional group interconversion to generate the canonical \( \beta \)-ketoacyl functional group. This ketide is fragmented, generating a synthon equivalent to malonyl-CoA and a diketide. This diketide undergoes a functional group interconversion to generate another \( \beta \)-ketoacyl functional group and then is transformed into two synthons, one equivalent to malonyl-CoA and one to acetyl-CoA.

Thus, the PKS for delta-hexalactone production is made up of an acetyl-CoA specific loading module and two malonyl-CoA specific extension modules, where the first extension module catalyzes a partial reduction of the beta-ketoacyl to an alcohol, and the second catalyzes complete reduction of the beta-ketoacyl to a methylene. \textit{E. coli} produces both of the substrates necessary for these transformations. (Bennett et al., 2009) The architectures of the extension modules for these transformations are presented in Figure 3.3 and the complete enzyme architecture is presented in Figure 3.4. Using previously elucidated mechanisms for each catalytic domain, we generated a proposed mechanism (Figure 3.5).
Figure 3.3 | **Modules required to synthesize δ-hexalactone via a PKS.** Catalytic domains are abbreviated as follows: Acyltransferase (AT), Acyl-Carrier Protein (ACP), Ketosynthase (KS), Ketoreductase (KR), Dehydratase (DH), and Enoylreductase (ER).

**Trial Design: Ac-Mal_{OH}-Mal_{H}**

Figure 3.4 | **Complete architecture for a putative modular megasynthase chassis capable of producing δ-hexalactone from acetyl-CoA and malonyl-CoA.**
3.2.2 - Amino Acid Sequence

A literature review identified amino acid sequences of the requisite catalytic domains. With optimal linker sequences designed in Chapter 2 we set out to build and test our design (Figure 3.4). Amino acid designs for each design module part were merged to create the complete amino acid sequence.

3.2.3 - Nucleotide Sequence
As codon usage is implicated in proper protein folding (Angov et al., 2008), and modular megasynthases are notoriously difficult to express in *E. coli*, we sought to match codon usage rates from other modular megasynthases by codon harmonization.

Thus, we again analyzed our modular megasynthase database, calculating the codon usage rate for each gene (Figure 3.7). The median codon adaptation index (CAI) was 0.697.

![Histogram of codon adaptation indices (CAIs) for each gene in the database of putative modular megasynthases. Median codon usage for all PKSs in the database was 0.697.](image)

We also checked if there were any position dependent codon bias that needed to be taken into consideration for our design and found none (Figure 3.8). Therefore, we used a CAI of 0.697 to generate our nucleotide sequence.
Figure 3.7 | Codon usage normalized by position. Codon usages were calculated across each modular megasynthase gene, with position normalized by gene length. This should show any conserved patterns in codon usage (i.e. is codon usage consistently higher or lower in specific positions).

The synthesized gene construct was then placed under an IPTG inducible T7 promoter. In addition, modular megasynthases also require a 4'-phosphopantetheinyl transferase (PPTase) for phosphopantetheinylation of ACPs, and it has been shown that expression of a heterologous PPTase is necessary for functional PKS expression in *E. coli.* (B. A. Pfeifer et al., 2001) We sought to produce our target strain in a single step with a single step modification of a wild type *E. coli* strain, so we also included IPTG inducible *sfp* for post-translational phosphopantetheinylation in the design of the complete construct, both expressed from an IPTG-inducible T7 promoter (Figure 3.9). (B. A. Pfeifer et al., 2001) The complete construct was 14,126 base pairs in length.
3.2.4 - Synthesis and Assembly

Once we had our nucleotide sequence in hand we set out to select a synthesis and assembly approach that would balance cost and time. As discussed in the introduction, double-stranded DNA of effectively any length can be purchased. Chip-based oligos cost about $0.04 per base pair (up to 230 bp) so our construct would only cost ~$600 but would require almost 100 oligos. However, cloning these oligos into the complete construct would be untenable. On the other end of the spectrum, the complete construct could be synthesized, at a cost of $0.45 per base pair, ~$6400 for the whole construct with no additional cloning required. Ideally, in synthetic biology, each ‘part’ is synthesized separately and stored in a repository of some kind for future assembly as needed.

3.2.4.1 - Part Synthesis

To keep with the spirit of synthetic biology we chose to order each ‘part’ as a separate double-stranded fragment (gBlock). For modular megasynthases, parts could be defined at two levels, either at the catalytic domain and linker level or, at the module level. Initially, we chose to synthesize catalytic domains and linkers. As some of these parts are used repeatedly in our initial
construct, we figured this approach would save on cost while also limiting the total amount of cloning required.

Homology-based cloning techniques have arisen recently and are high-efficiency, scarless tools for DNA assembly. In general, these techniques require tracts of sequence homology between pieces. Thus, once we had received the gBlocks, we used primers and PCR to install the appropriate homology arms, allowing for scarless homology-based cloning techniques (Figure 3.10).

![Diagram](image.png)

**Figure 3.9** | Approach to assembling modular megasynthases from gBlock ‘parts.’

3.2.4.1.1 - Gibson Assembly of Parts
To assemble the complete construct from catalytic domain and linker gBlocks would require stitching 30 pieces together. The canonical method for homology-based cloning is Gibson Assembly, and Gibson Assembly has been proposed to be able to stitch numerous DNA pieces together in a single, one-hour, isothermal reaction. (Gibson et al., 2009) Specifically, commercially available Gibson kits have been proposed to enable assembly of up to 15 pieces in a single reaction. Thus, Gibson Assembly seemed to be the ideal tool for part assembly. As Gibson is limited to only 15-piece assemblies, we chose to iteratively assemble our construct by first making the modules from gBlocks and then stitching the modules together to complete the construct.

We began by attempting linear assembly of modules in a single experiment. Unfortunately, these experiments were unsuccessful, even after tuning a variety of parameters such as quantity of DNA, ratio of DNA, even kit type. Thus, we tried iterative linear assembly of modules from an even lower level by performing Gibson Assembly on fewer pieces (e.g. 2 or 3) and then iteratively assembling these sub-constructs into modules and then modules into the complete construct. This also failed. The linear 2-3-piece assemblies were successful (Figure 3.11), but we were never able to successfully achieve the next step of stitching these together into larger 4-6-piece assemblies.

![Figure 3.10](image)

**Figure 3.10 | Gel Electrophoresis of Gibson assemblies.** Ladder (left lane). Failed large assembly (middle lane). Successful 2-piece assembly of KS-AT (right lane).
This was largely due to an inability to successfully perform PCR. Gibson Assembly yields relatively little DNA, certainly not enough to then perform additional Gibson Assemblies, so each successful reaction needs to be separated by gel electrophoresis, gel-extracted, and then amplified. The 2-3-piece assemblies were separated and gel-extracted but never successfully amplified for additional cloning. This could have been due to a few factors. First, the chaotrope Guanidine-HCl from the gel extraction procedure is a known PCR inhibitor, so carry over from the gel extraction, combined with relatively low initial template concentrations could have been a cause of the failed PCR. In addition, the mechanism of Gibson Assembly involves an exonuclease that chews back 5’ ends. As the reaction is isothermal, this chewback occurs continuously, so after a linear assembly is successful, the 5’ ends of the new linear dsDNA will continue to be chewed back. If the chewback is significant enough, entire priming sites could be removed, again preventing PCR.

After attempting linear Gibson Assembly, the next method to try was a circular Gibson, in which the constructs we were trying to assemble would be stitched into a plasmid backbone, that was then cloned. We chose not to attempt this method because of time constraints. Each experiment is dramatically longer, after the Gibson Assembly is complete it is transformed into *E. coli*, the transformation is recovered and plated, and the plates take ~12 hours to reach visible colonies. These colonies then have to be checked to identify correct assemblies, the colonies have to be outgrown overnight, miniprepped, and then the assemblies retrieved via PCR. For iterative assembly of 30 pieces, this would likely take weeks or longer.

3.2.4.1.2 - Yeast Transformation Assisted Recombination (TAR) Assembly of parts
Yeast Transformation Assisted Recombination (TAR) is another homology-based cloning technique that relies on the native recombination systems in yeast and is commonly used to assemble parts in synthetic biology studies. (Gibson, 2009; Shao, Zhao, & Zhao, 2009) Briefly, DNA (single stranded or double stranded) to be assembled is transformed into yeast along with a linear plasmid backbone. Yeast’s native recombination system assembles the DNA and linear plasmid backbone into a selectable circular plasmid. It is very similar in nature to Gibson Assembly but is more robust, allowing many more pieces of DNA to be assembled in a larger range of concentrations and requires a yeast transformation. Although colony formation takes longer in yeast than *E. coli* the robustness of the system more than covers the trade-off. Given our failures in using Gibson Assembly, we decided to attempt yeast TAR to assemble our entire construct.

Our attempt to assemble the complete construct in yeast failed, yielding few colonies that were all background (circularized vector). Again, as with our approach to Gibson assembly, we decided to assemble the modules next. This experiment was successful. We were able to assemble the Ac loading module as well as the Mal₅ and Mal₆ extension modules (Figure 3.12).

![Image](image.png)

**Figure 3.11 | Gel electrophoresis of junction PCRs.** A) Junction PCRs for Mal₆OH extension module. B) Junction PCRs for Mal₃ extension module. The KS-AT junction did not really work but, a faint band of the complete module was at the right size (6.0kb).

The next step was to then amplify these assemblies and use them in another yeast TAR experiment to complete our construct. This experiment was also a failure. Presented in Figure 3.10
is the gel electrophoresis of the junction PCRs we designed to confirm successful assemblies. Briefly, we designed PCRs that would amplify each junction in the assembly, allowing us to confirm that two pieces were successfully connected. In order to (likely) have the correct assembly, a colony would need to have positive results for every junction. In Figure 3.13 we can see that the colony presented has only 5 of 7 positive junction PCRs. This pattern occurred in almost 100% of colonies selected. Initially, as we did not have a positive control for these PCRs, we decided to troubleshoot whether this was a PCR failure or, a failure of the assembly process.

![Figure 3.12 | Yeast assembly of complete construct. All junctions are present except for internal junctions joining the two extension modules.](image)

To do this, we miniprepped the yeast plasmid. As we had used a shuttle vector containing replication origins from both *S. cerevisiae* and *E. coli*, we were able to then transform this plasmid into *E. coli*. As the replication origin in *E. coli* is much higher copy than the replication origin in *S. cerevisiae*, we were able to isolate enough DNA to sequence. The sequencing results showed an interesting result, we were only getting a single module in our assemblies, rather than both. This is why, in our junction PCRs we saw the first extension module connected to the loading module and the second module connected to the thioesterase but none of the internal junctions. Our
working hypothesis is that because the extension modules were assembled from parts with the same nucleotide sequence, they had a large amount of internal homology, allowing them to recombine with each other in a kind of crossover event (Figure 3.14). This may have occurred because we did not provide enough of the linker DNA to connect the two extension modules or due to some kind of secondary recombination that deleted this region of our construct after the initial assembly.

![Diagram of Proposed mechanism for cross-over.](image)

**Figure 3.13 | Proposed mechanism for cross-over.** Internal homology between the two extension domains is proposed to yield the internal junction deletion.

3.2.4.2 - Module Synthesis

Given that assembling our complete construct from individual parts was a failure, we chose a new approach to assembly. We decided to have complete modules synthesized independently, allowing us to perform yeast assembly of the final construct without the concern of internal
homology leading to crossover events that delete regions of our construct. This drives up the cost of synthesis but still allows us to treat each module as an independent ‘part’.

3.2.4.2.1 - Yeast Transformation Assisted Recombination (TAR) Assembly of modules

This approach was successful. The construct was synthesized in pieces and assembled with two rounds of yeast TAR (Figure 3.15). Again, we performed junction PCRs after each round to confirm the construct was appropriately assembled (Figure 3.15). At each step in the process, all junctions were present as desired.

Figure 3.14 | Yeast assembly of synthesized modules. A) Constructs synthesized by GenScript. B) Backbone yeast assembly. The final two lanes in the gel are successful assemblies. C) Integration of extension modules into backbone. All junctions are present after junction PCR.

3.2.4.3 - Transformation of High-Copy Large Plasmids into E. coli
After assembly of the full construct was complete, we sought to sequence the plasmid as validation. For a construct of this size (>14kb) this requires a significant amount of miniprepped DNA, much more than the yield of a miniprep of a single copy yeast plasmid. In addition, genome integration of the construct requires a significant amount of linear DNA. If PCR is not feasible, this linear DNA must come from a restriction digested plasmid. This will be discussed further in the next section. Both of these considerations required cloning the shuttle vector containing the complete construct into *E. coli*.

Initially, we tried to transform kit-based yeast minipreps into electrocompetent cells we had prepared of E. cloni and DH10B strains of *E. coli*. These were unsuccessful. We hypothesized that improved transformation efficiency would be necessary, because our plasmid was so large (~19kb) and as vector size increases, transformation efficiency drops. (Hanahan, 1983)

We then tried commercial electrocompetent strains with transformation efficiencies in excess of 10^10. Again, this was unsuccessful. This time we hypothesized that perhaps the amount of DNA we were transforming was too small, so we scaled up to a yeast midiprep. This was also unsuccessful. As the pUC origin is very high copy (300-500 copies per cell), and the plasmid is so large, we hypothesized that this much excess DNA in the cell might be quite toxic. (Imanaka & Aiba, 1981) Thus, we set out to find a way to decrease pUC origin plasmid copy number without having to start the assembly process over from the beginning with a new shuttle vector containing a different *E. coli* origin.

The pUC origin of replication is a derivative of the ColE1 origin of replication but, it contains a point mutation in the replication primer, RNA II, and is devoid of the Rop protein. This point mutation decreases primer binding to RNA I and results in increased replication initiation and thus copy number. It has been shown that the effects of this mutation are temperature
dependent. At higher temperatures, the effect is magnified, yielding even higher copy numbers and at lower temperatures, the effect is minimized, yielding lower copy numbers. (Lin-Chao, Chen, & Wong, 1992) Thus, we tried transforming out construct into *E. coli* and then recovering the strain at 30°C in order to minimize copy number. This approach was successful, allowing us to miniprep large amounts of plasmid from *E. coli* for genome integration and sequencing.

3.2.4.4 - *E. coli* Strain Preparation and Genome Integration

Evidence suggests that genome integration of modular megasynthases yields better results in *E. coli* so we integrated our construct into the genome of BL21(dE3)STAR using a CRISPR-based method previously developed by our lab (Figure 3.16). (Bassalo et al., 2016; Wang & Pfeifer, 2008) This method requires a strain that is both recombinogenic due to the presence of the lambda proteins and contains the Cas9 protein. Thus, we transformed two plasmids into BL21(dE3)STAR: pSIM5 containing beta, exo, and gam as well as pX2-Cas9 (Figure 3.17). (Datta, Costantino, & Court, 2006)

**Figure 3.15** | Genome integration of construct. The complete yeast assembly is restriction digested and integrated into the *E. coli* genome. Again, all junction PCRs are successful.
With this strain prepared, we set out to integrate our construct into the genome. The genome integration method cited above showed much better results using a linear integration template rather than a circular one. That study used a PCR amplified template. Thus, we attempted both PCR amplification of our complete construct to generate the integration template, as well as transformation of the shuttle vector containing the complete construct as template. Neither of these were successful, as the construct was likely too long to successfully PCR amplify and the plasmid transformation efficiency was likely too low. Instead, we attempted to digest our plasmid at restriction sites directly flanking the construct. This approach enabled successful genome integration (Figure 3.16).

We then genome sequenced the integration locus to confirm the successful integration of the construct and to preclude accidental introduction of any deleterious mutations during the cloning process. We identified three point mutations and no insertions or deletions. One point mutation was synonymous and the missense mutations were A1912V and H1997Y.

3.2.6 - Proteomics

Next, we chose to perform label-free quantitative bottom-up shotgun proteomics to confirm appropriate expression of our target protein upon induction. In a single experiment, we were able
to detect expression of our target, as well as *sfp*, and monitor any other significant changes to the proteome when compared to a negative control that may suggest sub-optimal expression. Our results showed that the target protein was highly expressed but was not significantly higher than native proteins (Figure 3.17). This was as expected and desired as dramatically overexpressed heterologous multimodular proteins often require significant amounts of chaperones for folding which can result in chaperone titration and the concomitant loss of folding and function. (Baneyx & Mujacic, 2004) In addition, sfp was also well expressed (Figure 3.17).

![Graph showing expression levels of genes](image)

**Figure 3.17 | Label-free quantitative shotgun proteomics results.** The top-20 genes by expression-level are sorted and plotted. Both genes in the designed construct are well expressed.

**3.2.7 - Small Scale Production for Metabolomics**
Ideally, our engineered enzyme would be robust enough and have the catalytic efficiency to produce detectable amounts of metabolite in a short, small scale experiment. Thus, we began by testing our design at small scale. We had already confirmed expression (section 3.2.6) so these experiments were meant to identify conditions appropriate for d-hexalactone production.

We used the proposed mechanism for our enzyme (Figure 3.4) to generate the overall reaction formula:

\[
1 \text{ acetyl-CoA} + 2 \text{ malonyl-CoA} + 4 \text{ NADH} \rightarrow d\text{-hexalactone} + 3 \text{ CoA} + 4 \text{ NAD}^+ 
\]

Given this reaction and without modifying the kinetics of our designed enzyme, d-hexalactone production is dependent on the intracellular concentrations of acetyl-CoA, malonyl-CoA, and NADH.

Under standard conditions, malonyl-CoA concentrations appear to be limiting, as *E. coli* maintains a very low steady-state malonyl-CoA concentration, below that of both acetyl-CoA and NADH. (Bennett et al., 2009) Culture conditions have also been implicated in intracellular metabolite concentrations. For example, acetyl-CoA concentrations have been shown to be negatively affected when *E. coli* is grown on any carbon source besides glucose. (Takamura & Nomura, 1988) As another example, a significant excess of NADH (or other reducing equivalents such as NAD(P)H) appears to be produced when *E. coli* is fermented using glycerol as a carbon source. (Durnin et al., 2009; Murarka, Dharmadi, Yazdani, & Gonzalez, 2008)

There are numerous metabolic engineering approaches for maximizing these concentrations however, metabolic engineering can be arduous and time consuming. (Atsumi, Cann, et al., 2008; Atsumi, Hanai, & Liao, 2008; Davis, Solbiati, & Cronan, 2000; Singh, Cher
Soh, Hatzimanikatis, & Gill, 2011; Zha, Rubin-Pitel, Shao, & Zhao, 2009) Therefore, our experimental design tested culture conditions to maximize the concentrations of each of these metabolites.

3.2.7.1 - Growth and Induction Conditions

To test for the growth condition parameters that would yield the largest amount of d-hexalactone, we used an array of different conditions. We began with rich media, both complex (LB) and defined (EZ Rich). We used a few different carbon sources: glucose, glycerol, glucose and glycerol, and a no carbon source control. In addition, we decided to test modifications to the dissolved oxygen in the cultures, testing aerobic, microaerobic, and anaerobic conditions.

It has been shown that oftentimes under laboratory conditions, bacterial secondary metabolites are produced in the stationary phase which is hypothesized to be most similar to growth conditions in the natural environment. (Price-Whelan, Dietrich, & Newman, 2006) As polyketides such as d-hexalactone are secondary metabolites, we sought to induce production in the stationary phase. The stationary phase is defined as the period in which growth has slowed from a constant doubling rate towards a rate in which growth and death are balanced. In addition, as we were using the IPTG-inducible lac operator, we needed to avoid excess glucose in our cultures at induction because excess glucose causes catabolite repression. (Novy & Morris—Novagen, n.d.; Ramey, 2002) Using glucose as a carbon source allows for fast growth however, a significant risk with using just glucose is that once the cultures have exhausted their carbon source and are induced, they also begin to die which negatively impacts production. Glycerol does not have the same catabolite repression effects that glucose does so glycerol can be added to cultures
at higher concentration, allowing for slower but longer and more sustained growth. This longer and more sustained growth allows for a longer productive induction time. Over the course of these experiments, we monitored glucose levels as well as optical density. The growth curves for these conditions are presented in Figure 3.18.
Figure 3.18 | Growth curves for each expression condition.

Our initial glucose experiments started at 1% glucose. As is demonstrated in Figure __, the LB+1% glucose cultures reached the stationary phase after about 6 hours at OD 3.3 but, the cultures still retained between 0.6-0.7% glucose. This suggests that our strain was depleting
another nutrient in the medium or perhaps producing a toxic amount of acid before depleting glucose, and induction would be unsuccessful. For future experiments we lowered the initial glucose concentration to 0.2%. For EZ Rich Medium, a standard glucose supplementation is 2%. This EZ+2% glucose experiment had a similar result to that of the LB experiment, reaching the stationary phase after about 5 hours with an OD of 1.4 and 1.4% glucose remaining. Thus, for future EZ experiments, we lowered glucose to 0.5%.

The glycerol experiments in both LB and EZ medium both gently tapered to the stationary phase at ~8hrs and OD 1.4-1.5. As suspected, the stationary phase included continued growth, as glycerol metabolism does not have the kind of overflow metabolism producing acid that glucose metabolism does. (Vemuri, Altman, Sangurdekar, Khodursky, & Eiteman, 2006; Xu, Jahic, Blomsten, & Enfors, 1999)

The last condition we tried in both LB and EZ was a combination of glucose and glycerol as carbon source. This appeared to offer the best of both worlds, as glucose metabolism allowed for quick growth to the stationary phase, and the switch to glycerol consumption allows for sustained growth without catabolite repression. This is similar to autoinduction medium as described by Studier. (Studier, 2005) LB + 0.2% Glucose + 1% Glycerol cultures reached the stationary phase around 6.5 hrs at OD 1.7 but, continued sustained growth to OD >5 at ~15hrs. EZ + 1% Glucose + 1% Glycerol cultures reached the stationary phase around 6 hrs at OD 1.3 and also sustained growth up to OD >3 at ~15hrs.

With this data in hand, we repeated the experiments, only this time inducing d-hexalactone production as each culture transitioned to the stationary phase. After 24 hours of induction the pellets and supernatants were reserved for extraction and detection, as we hypothesized that the metabolite may either a) diffuse through the cell membrane or b) be excreted by the cell.
As described in the introduction, methods for metabolite extraction must be compatible with the chemical properties of the target molecule as well as the downstream detection approach. D-hexalactone is volatile, with a boiling point of 231°C, making it compatible with GC-MS. (Burdock, 2016) In addition, it is relatively nonpolar, making it difficult to separate using standard reverse phase liquid chromatography. The lack of functionality also likely makes it difficult to ionize in an LC-MS/MS ESI source, compared to an EI source in a GC-MS. (de Hoffmann & Stroobant, 2007) Thus, we decided to try GC-MS detection first.

We began with the reserved pellets and supernatants of LB cultures from section 3.2.7.1. GC-MS requires a volatile non-polar solvent. For the cell pellet, the cells must be lysed directly into a GC-MS compatible solvent or, lysed and then extracted into an appropriate solvent. As the supernatant is aqueous, the volatile metabolites must be extracted into an appropriate solvent. For the cell pellet, we began by lysing the cells into an aqueous lysis buffer, followed by an extraction into an Ethanol:Methyl tert-Butyl Ether (EtOH:MTBE) mixture. This solvent mixture is proposed to separate the lipidome from the metabolome, simplifying the matrix for detection. (Chen et al., 2013) In addition, we used smaller solvent volumes than culture volume, slightly concentrating the metabolites. The MTBE layer was collected for injection onto the GC-MS.

For the supernatant, direct liquid-liquid extraction is less appealing, as concentrating the extract is much more difficult; however, we had access to an inline Solid Phase Microextraction (SPME) GC-MS. In SPME, a fiber is coated with an extracting phase. In our application, the fiber
was used to sample the headspace of the supernatant for volatiles. We saturated the supernatant with NaCl to ‘salt-out’ the d-hexalactone, driving the gas-liquid equilibrium towards gas. (Sada, Morisue, & Yamaji, 1975) After the SPME fiber was introduced into the headspace, the sample was heated, evaporating the volatiles into the headspace where they adsorbed onto the SPME fiber. The SPME fiber was then desorbed into the GC-MS.

The GC-MS used in both experiments was the same. The retention time for the authentic standard was 8.8 min in SPME experiments. No d-hexalactone was detected in the supernatant (Figure 3.16) For liquid injections, the standard had a retention time of 5.8 min. A small peak was detected in the cell pellet extract at 5.8 min (Figure 3.17). The mass spectra of both the experimental and control samples at 5.8 min matched (Figure 3.18). The experimental sample had such a small amount of d-hexalactone that there is some column bleed contamination. The detection limit for our approach was 22ng per culture, and the limit of quantification was 75ng per culture.
Figure 3.19 | Solid-phase microextraction results. Solid-phase microextraction results for: δ-hexalactone standard (red), 20C induction (blue), 30C induction (pink), 37C induction (brown), and negative control (black). There is no δ-hexalactone present in any experimental samples.
Figure 3.20 | Gas chromatogram of liquid injection. MTBE pellet extract was injected, and a peak was present at 5.8 min, the same retention time as the authentic standard.

Figure 3.21 | Mass spectra of liquid injections. On the left is the mass spectra for the experimental sample at a retention time of 5.8 min. On the right is the mass spectra for the authentic standard. These spectra match, with a little bit of column bleed in the experimental sample (281.1 m/z)
With a detection method in hand, we set out to quantify \( \delta \)-hexalactone production in each of the culture conditions described in section 3.2.7.1. These conditions included replicates of the LB cultures used in the detection experiments described above. Unfortunately, \( \delta \)-hexalactone was not detected in any of these samples (Figure 3.19 and 3.20), including the LB replicates. This was evidence that our initial result identifying \( \delta \)-hexalactone may have been a false positive.

**Figure 3.22 | Extracted ion chromatograms (XICs) for LB culture conditions.** \( \delta \)-hexalactone positive control XICs (m/z 114 and 99) compared to negative control and different LB culture conditions. Each sample was run with three replicates and there is no \( \delta \)-hexalactone in any experimental sample.
To test this, we returned to our LB samples, retesting the identified extraction conditions as well as new extraction conditions. For the new extraction conditions, we modified the ratio of H2O:EtOH:MTBE. MTBE is partially soluble in water, so by increasing the ratio, MTBE would better partition out of the water. We also decided to try SPME on the cell pellets, lysing them into a water, saturated NaCl solution. Again, none of these experiments yielded d-hexalactone.

Figure 3.23 | Extracted ion chromatograms (XICs) for EZ Rich culture conditions. δ-hexalactone positive control XICs (m/z 114 and 99) compared to negative control and different EZ Rich culture conditions. Each sample was run with three replicates and there is no δ-hexalactone in any experimental sample.
Lastly, we tried LC-MS/MS with aqueous lysates. Unsurprisingly, LC-MS/MS detection limits were 216.9 ng per culture, about 10X that of GC-MS detection methods. The increased detection limit also meant that d-hexalactone was not detected in the experimental samples.

3.2.8 - Large Scale Production for Metabolomics

Given that all of our small-scale experiments failed, with our final experiment we set out to maximize total production. This meant scaling up, maximizing substrate concentrations, and using an extraction method that included a concentration step. In addition, for scale up we chose to use a fermentor, so we could control conditions and feed the culture over many hours, allowing dramatically more time for production as well as dramatically more catalyst (i.e. cells). We hypothesized that these conditions would give us the best chance of detection, as even if d-hexalactone were produced at extremely low levels, we should be able to detect it.

For example, the GC-MS detection limit identified in section 3.2.7.2 was 22ng per culture. Scaling from 10mL to 500mL of culture, growing to OD 50 instead of 2, and inducing for 96 hrs instead of 24 would represent a 5000X increase in production. This would imply that even if our small-scale cultures were producing at as low as the picogram scale, scale up should allow us to bring this production into detectable range.

3.2.8.1 - mat operon for Malonate Feeding

Before fermenting our strain, we needed a method to maximize intracellular substrate concentrations. As described in section 3.2.7, there are have been numerous metabolic engineering
approaches to increase malonyl-CoA and acetyl-CoA concentrations. These methods though, require significant amounts of genetic engineering and we were looking for a faster, simpler approach.

For heterologous secondary metabolite production in *E. coli*, it is fairly common for *E. coli* to lack the required substrates. Exogenous substrate feeding has proven an effective method for overcoming this. (Pacey et al., 1998; B. A. Pfeifer et al., 2001; Blaine A. Pfeifer, Wang, Walsh, & Khosla, 2003) A literature search failed to identify methods that exogenously feed either acetyl-CoA or malonyl-CoA, likely because both malonyl-CoA and acetyl-CoA are extremely expensive, ~$10,000 per gram. *E. coli* does contain a gene that produces acetyl-CoA from exogenous acetate, acetyl-CoA synthetase (Acs). (Brown, Jones-Mortimer, & Kornberg, 1977; Kumari et al., 2000) Acs is cotranscribed with yjcG, an acetate permease. (Gimenez, Nuñez, Badia, Aguilar, & Baldoma, 2003) Finally, acetyl-CoA can be converted to malonyl-CoA by acetyl-CoA carboxylase (Acc). Overexpression of Acc has been shown to increase malonyl-CoA levels in *E. coli*. (Davis et al., 2000; Zha et al., 2009) So one approach to generating increased acetyl-CoA and malonyl-CoA levels would be overexpression of yjcG, Acs, and Acs, and exogenous feeding of acetate. This approach has the downside that acetate is toxic to *E. coli* so only a limited amount could be fed. (Sandoval, Mills, Zhang, & Gill, 2011)

We chose another route to producing acetyl-CoA and malonyl-CoA via exogenous feeding. Previously, an operon in the legume root-associated bacterium *Rhizobium trifolii*, matABC, was identified. This operon allows for the import of exogenous malonate into the cell (matC), conversion of malonate to malonyl-CoA (matB), and malonyl-CoA conversion to acetyl-CoA (matA) (Figure 3.21). (An & Kim, 1998) Introduction of these genes into *E. coli* has been shown to allow malonyl-CoA production at sufficient levels to allow growth in the absence of functional
Acc, as well as allowing exogenous feeding of malonate for polyketide production. (Lombó et al., 2001)

Figure 3.24 | Mechanism of the mat operon. MatC imports malonate into the cell, matB converts malonate into malonyl-CoA, and matA converts malonyl-CoA to acetyl-CoA. Malonyl-CoA and acetyl-CoA are then converted to δ-hexalactone by our design.

To optimize these genes for expression in *E. coli*, we used the Operon Calculator as reported by the Salis Lab to generate an operon sequence with desired promoter, terminator, and relative gene expression levels. (Tian & Salis, 2015) As malonyl-CoA is the limiting substrate we designed the operon to express matB and matC at twice the level of matA, allowing for accumulation of malonyl-CoA. We had the operon synthesized and cloned it into our strain.

To test that the operon was functioning as desired, we needed to demonstrate that expression of these three genes generated increased malonyl-CoA levels upon malonate feeding. To do this we used a previously reported flaviolin reporter system. (Kuščer et al., 2007; Magdevska et al., 2010) Flaviolin (2,5,7-trihydroxy-1,4-naphtoquinone) is a red polyketide formed by the
spontaneous oxidation of 1,3,6,8-tetrahydroxynaphthalene, which is synthesized from 5 units of malonyl-CoA by the gene RppA from *Streptomyces griseus* (Figure 3.22). (Funa, Ohnishi, Ebizuka, & Horinouchi, 2002; Ueda, Kim, Beppu, & Horinouchi, 1995) Flaviolin can be used as a reporter for malonyl-CoA levels, as the redness of the culture will be a function of malonyl-CoA concentration, assuming the malonyl-CoA concentration is below saturation for the enzyme and the enzyme kinetics are held constant. (Tarasava, Liu, Garst, & Gill, 2018)

**Figure 3.25 | Flaviolin biosynthesis.** RppA converts 5 units of malonyl-CoA to 1 unit of THN. THN then spontaneously oxidizes into Flaviolin.

We performed two experiments to test the efficacy of the synthesized matABC operon. First, we tested whether flaviolin levels changed with varying concentrations of malonate feeding in a strain carrying our synthesized operon. Figure 3.23, shows that increasing levels of malonate feed yield increased levels of flaviolin in this strain.
Figure 3.26 | Flaviolin production. In the presence of our matABC construct, Flaviolin production as measured by absorbance at 500nm increases with exogenously fed malonate to an exogenous concentration of 10mM.

Although, at malonate concentrations above 10mM flaviolin levels begin to drop. This is likely due to some kind of inhibitory effect. For example, high levels of malonate could very well be toxic to cells, as the pH of the growth medium drops with malonate addition. Second, we compared the effects of malonate feeding on flaviolin production in the strain containing the matABC operon and a negative control to confirm that the mat operon is causing the effect we identified in the initial experiment. The results of this experiment are presented in figure 3.24. In the negative control strain, malonate feeding has no impact on flaviolin levels. On the other hand, as identified in our initial experiment, malonate feeding resulted in increased levels of flaviolin production in the strain containing the synthesized matABC operon. These results suggest that the matABC operon works as hoped and provides an ideal method for maximizing acetyl-CoA and malonyl-CoA concentration by exogenous feeding of malonate.
Figure 3.27 | Flaviolin production in negative control. A comparison of Flaviolin production between strains with (orange) and without (blue) the mat operon. The mat operon appears to yield increased Flaviolin when fed malonate exogenously, while the negative control does not.

3.2.8.2 - Fermentation

The next step was to perform fermentations of our strain carrying the matABC operon. To do this, we worked from a previously reported method demonstrated to be highly effective for polyketide production in *E. coli* (Figure 3.28). (B. Pfeifer, Hu, Licari, & Khosla, 2002)
Figure 3.28 | Fermentation of Ac-Mal$_{\text{OH}}$-Mal$_{\text{H}}$.

The fermentation was induced with 1mM IPTG as the culture transitioned to the stationary phase after consumption of 5g/L glucose, at OD 7.4. A glucose feed was started with glucose concentration in the fermentation maintained below 0.5 g/L so as to avoid overflow metabolism. In addition, malonate was added to a concentration of 10mM. Malonate concentration was monitored via HPLC (Figure 3.29).
Figure 3.29 | Malonate HPLC traces. Fermentation medium with 10mM malonate has a peak at 9.5min, while fermentation medium with no additional malonate has no peak at 9.5 min.

The authentic malonate standard had a retention time of 9.5 minutes. This peak was present upon initial malonate feeding (Figure 3.30).

Figure 3.30 | HPLC trace post fermentation induction. The malonate peak at 9.5 min is present.

After 12 hours, a small amount of malonate remained (Figure 3.31) and after 18 hours there was no malonate remaining (Figure 3.32).
We chose to feed malonate to a concentration of 10mM every 16 hours. The fermentation was continued for 84 hours after induction, to an OD of 55 and then harvested.

3.2.8.3 - Metabolite Extraction and Detection

Extracts of the target strain, a negative control, and the authentic standard were each analyzed via GC-MS. The δελ.τα-hexalactone standard had a retention time of 3.66 minutes. The
negative control had no peak at 3.66 minutes, whereas there was a peak present in the extract of our engineered strain. The chromatograms from each sample are overlaid in Figure 3.33.

Figure 3.33 | Fermentation extract gas chromatogram. Gas chromatogram of negative control (pink), experimental fermentation (black), and authentic δ-hexalactone standard (blue). The authentic standard has a retention time of 3.65 min. There is a small peak at this retention time in the authentic standard, and none in the negative control.

The mass spectrum for the ™-hexalactone standard (Figure 3.34) and the 3.66-minute retention time peak from our strain (Figure 3.35) match.
Figure 3.34 | δ-hexalactone mass spectrum. Mass spectrum for δ-hexalactone authentic standard (3.65min retention time).
Figure 3.37 | Fermentation extract mass spectrum. Mass spectrum of the fermentation extract at a retention time of 3.65 min. This is a perfect match to the δ-hexalactone authentic standard.

In addition, a library search against the NIST MS database identified the compound at 3.66 minutes in our sample as δελτα-α-hexalactone, confirming the function of our design.

3.3 - Discussion

We have demonstrated the ability to successfully design, assemble, and express a functional megasynthase enzyme from the bottom-up. This was enabled by our novel approach to
modular megasynthase modularity that treats catalytic domain linkers as modular parts, as well as the catalytic domains themselves.

In chapter 2, we reported a computational design tool that yields putative catalytic domain linker ‘parts’. These linkers, when combined with catalytic domains, enabled de novo design of multiple modules and a complete chimeric modular megasynthase chassis, suggesting that the linkers are context-independent. At the module level, the same linker amino acid sequence was used in multiple modules; for example, the KS-AT linker amino acid sequence was the same in both extension modules and both modules functioned as intended.

There is further validation that can be done for the designed megasynthase beyond qualitatively proving that the enzyme is producing d-hexalactone. Specifically, the production of the strain could be quantified. This requires identifying an appropriate internal standard to control for extraction efficiency (C13 labelled d-hexalactone would be ideal for this purpose). In addition, it would likely take five to seven repeats. Further, the kinetics of the enzyme itself could be determined.

From these parameters, additional optimization could be done through an approach like directed evolution (to improve the enzyme kinetics) or metabolic engineering (to improve the substrate levels) to generate a strain capable of producing d-hexalactone at an industrially relevant level. Unfortunately, most of these experiments would be much more feasible if the experimental conditions for d-hexalactone production were scalable. This requires identifying the minimal conditions for detectable d-hexalactone production such as minimal volume, induction time, malonate feeding, etc. There are enough parameters to optimize/minimize that a structured, design of experiment style factorial design would be needed to map a response surface. Ideally, production could be decreased to microtiter 96-well format or even higher throughput.
Of more relevance to this work is validation that this approach can be expanded beyond a single modular megasynthase. This is where we chose to focus our efforts, and the extensibility of the approach is investigated in chapter 4.

3.4 - Methods

3.4.1 - Amino acid designs

For a given modular megasynthase chassis, requisite catalytic domain amino acid sequences were identified in the literature. Modules were designed using canonical design rules for PKSs, NRPSs, and PKS-NRPS hybrids. Catalytic domain amino acid sequences were combined using appropriate linker amino acid sequences to yield complete module amino acid sequences. Module amino acid sequences were combined using appropriate linker amino acid sequences to yield the amino acid sequence of the complete modular megasynthase chassis.

3.4.2 - Codon Harmonization

To generate a nucleotide sequence for each chassis we set out to match codon usage rates of putative modular megasynthases. To do this we calculated codon usage rates for every putative modular megasynthase identified by antiSMASH 2.0 as described in ‘Genome Mining’. Briefly, for each genome, a codon usage table was generated using BioPython’s CodonUsage Module in which every CDS in the GenBank file was used to generate a codon usage table for the given genome. This codon usage table was then used to calculate codon usage rate for each putative
modular megasynthase identified by antiSMASH in the genome. The median codon usage of the entire database of modular megasynthases (0.7) was used to design nucleotide sequences.

3.4.3 - Other Design Considerations

As we set out to use a standard and easily accessible strain of *E. coli* (in this case, BL21(dE3)STAR), we had to include regulators, promoters, ribosome binding sites (RBSs), etc. that were compatible with this strain. We chose the IPTG inducible promoter and RBS from the pET system. In addition, we included the PPTase sfp from *B. subtilis* in the construct to minimize cloning so we placed this gene downstream of the modular megasynthase gene, with an insulating region in between, also under the IPTG inducible promoter and RBS from the pET system. For genome integration, both the 5’ and 3’ end of the complete construct included 100bp homology arms to the targeted integration site, between base pairs 3,876,433 and 3,876,434.

3.4.4 - DNA Synthesis

The complete construct was codon harmonized with a codon usage rate of 0.7 by GenScript. We then broke this construct into the synthesized pieces. Initially, we had each part (ie catalytic domain, linker, homology arm, or sfp) synthesized as a gBlock by IDT. We also had five larger, module-size parts synthesized by GenScript. The first piece synthesized by GenScript contained the loading module and TE domain consecutively. The second and third piece synthesized by GenScript were each of the extension modules. The insulator, sfp, and 3’ homology arm were synthesized by IDT as a gBlock with homology to the TE domain and the shuttle vector.
pRS416. The 5’ homology arm was also synthesized by IDT as a gBlock with homology to the loading module and pRS416.

3.4.5 - Gibson Assembly

Homology arms for Gibson Assembly were installed with primers via PCR amplification of gBlocks. Gibson Assemblies were performed using the Gibson Assembly HiFi 1-Step and Gibson Assembly Ultra Kits (Synthetic Genomics, Inc). The protocols for each kit were followed as outlined by SGI.

3.4.6 - Yeast Assembly

To assemble the complete construct we performed two rounds of yeast transformation assisted recombination (TAR). In the first round we assembled the 5’ homology arm gBlock, the PCR linearized loading module and TE domain construct, the sfp gBlock, and the PCR linearized pRS416 shuttle vector.

For yeast TAR we prepared an overnight of *S. cerevisiae* BY4709 with 25mL YEPD @ 30C and shaking @ 225 rpm. The overnight was diluted to 100mL in YEPD @ OD 0.2 and grown at 30C to OD 0.4-0.6 (~4 hrs). The culture was spun down at 3000g for 3 min (room temp), resuspended in 25mL dH2O (room temp), spun down again at 3000g for 3 min (room temp), and finally resuspended in 500uL mixture of 1X TE and 1X LiAc.

10uL of carrier DNA (10mg/mL sheared, denatured salmon sperm DNA in 1X TE) boiled for 15 min and then cooled to 0C on ice was added to equimolar mixture of DNA pieces. This
mixture was combined with 100uL LiAc yeast resuspension and vortexed. 600uL of a 40% PEG-4000, 1X TE, 1X LiAc solution was added. This mixture was vortexed for 20-30s and then incubated @ 30C for 1hr. 70uL of DMSO was then added to the transformation, incubated at 42C for 15 min, chilled on ice for 5-10 min, 600uL of 1X TE was added. This mixture was mixed by inverting and then centrifuged @ max for 10s. The supernatant was removed and the pellet resuspended in ~50uL 1X TE. The transformation was plated on YNB (-URA) and incubated at 30C until colonies were visible (~3.5 days). PCR was used to identify a construct with the appropriate assembly.

For the second round, we linearized the construct produced in the first round with primers that primed at the interface between the loading module and the thioesterase domain. The extension modules were also PCR amplified. The transformation was completed as described above. PCR across each junction in the assembly was used to confirm.

3.4.7 - Shuttle Vector Isolation from Yeast

3.4.7.1 - Yeast Miniprep

All yeast minipreps were performed with the Zymoprep Yeast Plasmid Miniprep II Kit (Zymo Research) following the provided protocols.

3.4.7.2 - Yeast Midiprep
Isolated yeast colonies are grown in 40mL YEPD for 24hrs @ 30C, shaking at 225rpm. The culture is then harvested by centrifuging at 5000rpm for 5 min. The pellet is resuspended in 3mL 0.9M Sorbitol, 0.1M Na:EDTA, pH 7.5. To this solution, 50uL zymolyase (5U/uL, Zymo Research) is added. The mixture is then incubated at 37C for 1 hour, and then spun down. The pellet is resuspended in 1mL P1 buffer (Qiagen). This solution is then divided into 4 250uL aliquots. To each aliquot, 250uL P2 buffer (Qiagen) is added. These mixtures are then incubated at 65C for 5 min, and then chilled on wet ice for 5 min. Finally, 350uL N3 buffer (Qiagen) is added. From the addition of N3 on, the Qiaquick Miniprep (Qiagen) protocol is followed. The purified plasmid is eluted in 50uL EB buffer (Qiagen).

3.4.8 - Shuttle Vector Transformation into E. coli

To transform the shuttle vector into E. coli we began by transforming 5uL of vector isolated via yeast miniprep into electrocompetent cells we prepared from lab stocks (E. cloni and DH10B).

To prepare electrocompetent cells, we grew a 2mL overnight culture in LB at 37C, shaking at 225rpm. The overnight was then diluted 1:100 in 50mL LB and grown to the mid-log phase (OD 0.5) at 37C and shaking at 225rpm. This culture was then chilled on ice for 15min and then spun down at 6000g for 5min at 3C. The pellet was resuspended in 50mL ice cold ddH2O, and spun down at the same conditions. This was repeated twice more, with 25mL and 12.5mL of ice cold ddH2O, respectively.

Commercial competent cells used were electrocompetent E. cloni 10G Supreme (Lucigen) and MegaX DH10B T1+ (Thermo Fisher).
Transformations were recovered in 1mL LB for 1hr at either 37C or 30C and then plated on LB agar plates supplemented with 100ug/mL carb. Plates were incubated for 24 hours at 37C or 30C. Colonies were tested via PCR of assembly junctions.

3.4.9 - Strain Preparation

As we intended to integrate our construct into the genome of *E. coli* BL21(dE3) STAR using our previously reported CRISPR-based method, we transformed Chemically Competent One Shot BL21(dE3)STAR from Thermo Fisher with the plasmid pX2-Cas9 (kanR, Addgene plasmid #85811) to introduce *S. pyogenes* Cas9. (Bassalo et al., 2016) The resulting strain was then also transformed with the plasmid pSim5 (cmR) to enable high efficiency homology-directed repair. (Datta et al., 2006) Strains requiring malonyl-CoA as a substrate were also transformed with a plasmid containing the matABC operon reported in Lombo et al. (Lombó et al., 2001)

3.4.8 - Genome Integration

Genome integration was performed as previously described. (Bassalo et al., 2016) Constructs were prepared for genome integration by restriction digestion of 1ug plasmid DNA with restriction enzymes directly flanking each homology arm. The restriction digestion was then concentrated using a 30,000MWCO Ultra-0.5mL Amicon Centrifugal Filter. The gRNA plasmid (ampR, Addgene #71656) targeting the genomic integration locus was previously described. (Bassalo et al., 2016) 100ng of miniprepped gRNA and 800ng linear integration template were used in the transformation.
3.4.9 - Genome Sequencing

Each genome integration was confirmed by genome sequencing. The integration locus was PCR amplified in overlapping pieces. Amplicons were prepared for sequencing with Nextera XT DNA Library Prep Kit. Samples were then sequenced on an Illumina MiSeq using a 150-cycle paired-end kit. Resulting FASTQ files were analyzed using Geneious v10.1.3. Briefly, FASTQ files were aligned to the *E. coli* BL21(dE3) genome sequence using the Geneious mapping algorithm in High Sensitivity/Medium mode with default settings.

3.4.10 - Label-free Shotgun Proteomics

Overnights of sequence verified strains were grown in 2mL LB supplemented with 100µg/mL carbenicillin at 37°C and shaking at 225rpm. New 15mL cultures were inoculated 1:100 with the saturated overnight cultures and grown to mid-log phase. The temperature was adjusted to 25°C and the culture was induced with 1mM IPTG. Pellets were harvested after 18 hours of induction.

Samples for LC-MS/MS analysis were prepared as described previously,(Kulak, Pichler, Paron, Nagaraj, & Mann, 2014) with minor modification. Briefly, *E. coli* cells were resuspended in a lysis buffer (6 M Guanidine HCl, 10 mM TCEP, 40 mM chloroacetamide, 100 mM Tris, pH 8.5), and boiled at 95 °C for 15 min. Cell lysate was, then sonicated and clarified by centrifugation at 13k rpm for 10 min. Protein concentration was determined using tryptophan fluorescence at 285nm/320 nm with a serial dilution of BSA as a standard. Seven microgram of each sample was
diluted 20-fold in a dilution buffer (5% acetonitrile, 50 mM Tris, pH 8.5) and proteolyzed using 0.25 µg of trypsin for 16 hrs at 37 °C. The resulting tryptic peptides were acidified using 1% (final) TFA, loaded onto in-house StageTips (3M SDB-RPS membrane), washed three times with 0.2% TFA. (Rappsilber, Mann, & Ishihama, 2007) The peptides were eluted with 70 µL of BufferX (5 mM ammonium hydroxide, 80% acetonitrile) and dried using vacuum centrifugation. Peptides were reconstituted in Buffer A (0.1% formic acid in water), 10% of which were subjected to LC-MS/MS analysis.

Aliquots (5 µl) of the resulting tryptic peptides were loaded directly onto a Waters nanoACQUITY UPLC BEH C18 column (130 Å, 1.7 µm × 75 µm × 250 mm) equilibrated with 0.1% formic acid/3% acetonitrile/water for UPLC-MS/MS analysis. The mobile phase A was 0.1% formic acid/water, phase B was 0.1% formic acid/acetonitrile and elution was at 0.3 µl/min using gradients of 3 to 8% B (0-5 min) and 8 to 32% B (5-123 min). MS/MS was performed on a LTQ Orbitrap Velos mass spectrometer, scanning precursor ions between 300 and 1800 m/z (1 × 10^6 ions, 60,000 resolution) and selecting the 20 most intense ions for MS/MS with 180s dynamic exclusion, 10 ppm exclusion width, repeat count = 1, and 30 s repeat duration. Ions with unassigned charge state and MH+1 were excluded from the MS/MS. Maximal ion injection times were 500 ms for FT (one microscan) and 250 ms for LTQ, and the AGC was 1 × 10^4. The normalized collision energy was 35% with activation Q 0.25 for 10 ms.

MaxQuant/Andromeda (version 1.5.2.8) was used to process raw files from LTQ-orbitrap, and search the peak lists against database consist of Uniprot E. coli K12 strain proteome (downloaded at 02/02/2016) plus sequence for engineered proteins (total 4,318 entries). The search allowed trypsin specificity with maximum two missed-cleavage and set carbamidomethyl modification on cysteine as a fixed modification and protein N-terminal acetylation and oxidation.
on methionine as variable modifications. MaxQuant used 4.5 ppm main search tolerance for precursor ions, 0.5 Da MS/MS match tolerance, searching top 8 peaks per 100 Da. False discovery rates for both protein and peptide were 0.01 with minimum seven amino acid peptide length. For quantification, “LFQ Intensity” for a protein group reported by MaxQuant were used.

3.4.11 - Small Scale Cultures

For small scale cultures, we prepared a 2mL culture of the sequence verified strain in LB Broth (Teknova, Hollister, CA) supplemented with 100ug/mL carbenicillin and grew overnight at 37°C, shaking at 245rpm. The overnight was diluted to OD 0.01 in 15mL media supplemented with 100ug/mL carbenicillin. The media bases used were LB Broth (Teknova, Hollister, CA) and EZ Rich Medium (Teknova, Hollister, CA). Concentrated stocks of 200g/L glucose and 500g/L glycerol in ddH2O. The conditions used with the LB media base were (w/v): no additional carbon source, 0.2% glucose, 1% glucose, 1% glycerol, and 0.2% glucose + 1% glycerol. The conditions used with the EZ Rich media base were (w/v): no additional carbon source, 1% glucose, 2% glucose, 2% glycerol, and 1% glucose + 1% glycerol.

Cultures for d-hexalactone production and detection were cooled to 25°C after reaching the stationary phase and induced with 1:1000 dilution of a 1M stock of sterile filtered IPTG (Gold Biotechnology, St. Louis, MO) in ddH2O. Induced cultures were harvested 24 hours post-induction and centrifuged. The pellets and supernatants were reserved for extraction and detection.

3.4.12 - Metabolite Extractions
3.4.12.1 - 3:3:1 (H₂O:EtOH:MTBE) Pellet Extraction for GC-MS

The thawed pellets were mixed with 100 µL of water, and then sonicated for 45 min at 4°C. The slurry was transferred to a glass vial with additional 200 µL water, and then mixed with 300 µL of ethanol and vortexed for 15 min. MTBE (100 µL) was added and then vortexed for another 5 min. The mixture was centrifuged and the clear solvent layer (100 µL) was transferred to autosampler inserts. The remaining materials (pellet solids, water, and organic solvents) were transferred to SPME autosampler vials with the aid of 1 mL of water, and then saturated with NaCl.

3.4.12.2 - SPME Supernatant Extraction for GC-MS

The supernatant (7 mL) was transferred into SPME autosampler vials and saturated with 30% (w/v) NaCl.

3.4.12.2 - 3:1:1 (H₂O:EtOH:MTBE) Pellet Extraction for GC-MS

The thawed pellets were mixed with 150 µL of water and 300uL NaCl saturated H₂O, transferred to a 1.5mL centrifuge tube and mixed with 150uL EtOH, and then sonicated with a Qsonica Q55 at 10% amplitude for 6 cycles, 10s on and 30s off. The slurry was vortexed for 30s transferred to a 2mL glass vial with 150uL µL MTBE, briefly and gently vortexed. The mixture was shaken at 600rpm for 1hr at 4C. The mixture was centrifuged at 3000g for 30min at 4C. The clear solvent layer (100 µL) was transferred to autosampler inserts.
3.4.12.3 - SPME Pellet Extraction for GC-MS

The thawed pellets were mixed with 1000 µL of NaCl saturated H₂O, transferred to a 1.5mL centrifuge tube, and then sonicated with a Qsonica Q55 at 10% amplitude for 9 cycles, 10s on and 30s off. The slurry was vortexed for 30s transferred to a 2mL glass vial with 150µL µL MTBE, briefly and gently vortexed. The mixture was shaken at 800rpm for 15min at 4C. The mixture was centrifuged at 14000rpm for 15min at 4C. The slurry was transferred to an SPME vial.

3.4.12.4 - Methanol Pellet Extraction for LC-MS/MS

To the frozen pellets, 1 mL of methanol was added. Sample was left at 4°C to thaw, then briefly vortexed and sonicated for 8 min using Qsonica Sonicator (50-amp, pulse on 30 s, pulse off 30 s). After that, samples were vigorously shaken for 1 hr at 4°C, and then centrifuged at 3720 g for 10 min. The supernatant was recovered.

3.4.13 - GC-MS

3.4.13.1 - SPME

Samples were incubated at 60°C for 10 min. Volatile compounds were extracted by a 75 µm CAR/PDMS autosampler fiber at 60°C for 20 min with agitation, and then desorbed at 260°C for 3 min into a TG-WAXMS column (30 m x 0.25 mm x 0.25 µm, Thermo) in a Trace1310 GC
(Thermo) coupled to ISQ-LT Mass Spectrometer. The fiber was post-conditioned for 10 min at 280°C before the next injection. GC inlet was set at 260°C and in splitless mode. A constant flow rate of the carrier gas (He) was controlled at 1.2 mL/min. The oven initial temperature was 100°C and held for 1.5 min, and increased to 240°C at 10°C/min and held for 3 min. Detection was completed under electron impact mode, with a scan range of 30-350 amu and scan rate 5 scans/second. Transfer line and source temperatures were 255 and 230°C, respectively.

3.4.13.2 - Liquid injection

The solvent extracts of pellets (1 µL) was directly injected to the same GC-MS as described above. GC inlet was set at 260°C, and split ratio was 10:1. The oven temperature was programmed from 100°C to 250°C at 20°C/min and held for 1 min each at the start and end temperatures. Other instrumental parameters were the same as those of SPME method.

3.4.14 - LC-MS/MS

Five µL of extract was injected onto a Waters Acquity UPLC system in randomized order, with a pooled QC injection after every 6 sample injections. Metabolites were separated using a Waters Acquity UPLC CSH Phenyl Hexyl column (1.7 µM, 1.0 x 100 mm), using a gradient from solvent A (2mM ammonium hydrioxide, 0.1% formic acid) to solvent B (Acetonitrile, 0.1% formic acid). Injections were made in 100% A, held at 100% A for 1 min, ramped to 98% B over 12 minutes, held at 98% B for 3 minutes, and then returned to starting conditions over 0.05 minutes and allowed to re-equilibrate for 3.95 minutes, with a 200 µL/min constant flow rate. The column
and samples were held at 65 °C and 6 °C, respectively. The column eluent was infused into a Waters Xevo G2 Q-TOF-MS with an electrospray source in positive mode, scanning 50-2000 m/z at 0.2 seconds per scan, alternating between MS (6 V collision energy) and MSE mode (15-30 V ramp). Calibration was performed using sodium iodide with 1 ppm mass accuracy. The capillary voltage was held at 2200 V, source temp at 150 °C, and nitrogen desolvation temp at 350 °C with a flow rate of 800 L/hr.

For each sample, raw data files were converted to .cdf format, and matrix of molecular features as defined by retention time and mass (m/z) was generated using XCMS software in R [1] for feature detection and alignment. Raw peak areas were normalized to total ion signal in R, and outlier injections were detected based on total signal and PC1 of principle component analysis. Features were grouped using RAMClustR, which groups features into spectra based coelution and covariance across the full dataset, whereby spectra are used to determine the identity of observed compounds in the experiment. Targeted analysis of the data was performed by searching the XCMS results, both at the individual peak and peak group levels, for signals consistent with the accurate mass of the target analytes. PCA was conducted on mean-centered and pareto variance-scaled data using the pcaMethods package in R.

3.4.15 - matABC Operon Design, Cloning, and Validation

The amino acid sequences for matA, matB, and matC from *Rhizobium trifolii* were accessed via their NCBI Accession Numbers: AAC83456.1 (matA), AAC83455.1 (matB), and AAC83457.1 (matC). (An & Kim, 1998) These sequences were input into the Operon Calculator, along with the promoter J23100 (Registry of Standard Biological Parts #BBa_J23100), and T1
terminator (Registry of Standard Biological Parts #BBa_B001) to generate an operon design for *E. coli*. (Tian & Salis, 2015) The targeted translation rates were 25000, 50000, 50000, respectively. All other parameters were left as default. The output sequence was synthesized as two overlapping gBlocks by IDT. The pBR322 backbone was linearized via PCR, with homology arms to the operon introduced via primers: ggctttcgttttatctgttggctcgggaacctctgccattttcaggggattgtgtctca and tgaatgtacacttgaggagctcgttgacagctacgtcagacagcaaggtaatttctca. The PCR product was cleaned using a QIAQuick PCR Purification Kit (Qiagen), dpnI digested, and gel extracted with a QIAQuick Gel Extraction Kit (Qiagen). The plasmid backbone and gBlocks were cloned into *E. coli* electrocompetent cells (Lucigen) using a Gibson Assembly HiFi 1-step kit (Synthetic Genomics). The plasmid (pBR_matABC, tetR) was sequenced verified.

To test the function of pBR_matABC, we transformed it into a BL21(dE3)STAR strain of *E. coli* carrying a plasmid (p_rppA) with IPTG-inducible *rppA*. RppA produces 1 molecule of the polyketide flaviolin from 5 molecules of malonyl-CoA. The lambda$_{max}$ of flaviolin is 500nm. Thus, *rppA* acts as a reporter for malonyl-CoA concentration, with readout of absorbance at 500nm. As a negative control, we used the BL21(dE3)STAR p_rppA without matABC.

All experiments were performed in minimal fermentation medium: 1.5 g of KH$_2$PO$_4$, 4.34 g of K$_2$HPO$_4$, 0.4 g of (NH$_4$)$_2$SO$_4$, 150.5 mg of MgSO$_4$, 5 g of glucose, 1.25 ml of trace metal solution, and 1.25 ml of vitamin solution. The feed medium composition (per liter) was: 110 g of (NH$_4$)$_2$SO$_4$, 3.9 g of MgSO$_4$, 430 g of glucose, 10 ml of trace metal solution, and 10 ml of vitamin solution. The trace metals solution was composed of (per liter): 27 g of FeCl$_3$ 6H$_2$O, 2 g of ZnCl$_2$ 4H$_2$O, 2 g of CaCl$_2$ 6H$_2$O, 2 g of Na$_2$MoO$_4$ 2H$_2$O, 1.9 g of CuSO$_4$ 5H$_2$O, 0.5 g of H$_3$BO$_3$, and 100 ml of concentrated HCl. The vitamin solution was composed of (per liter) 0.42 g of
riboflavin, 5.4 g of pantothenic acid, 6 g of niacin, 1.4 g of pyridoxine, 0.06 g of biotin, and 0.04 g of folic acid.

We grew 2mL overnights of both strains in LB Broth supplemented with 10ug/mL tetracycline and 50ug/mL kanamycin (BL21(dE3)STAR p_rppA pBR_matABC) or 50ug/mL kanamycin (BL21(dE3)STAR p_rppA) at 37°C shaking at 225rpm. We centrifuged 1mL of the overnights and resuspended them in 1mL of fermentation medium. The resuspended overnights were then diluted 1000X in fermentation medium supplemented with 1mM IPTG, the appropriate antibiotics and varying concentrations of malonate (0mM, 7.5mM, 10mM, and 12.5mM). The cultures were grown at 37°C, shaking at 215rpm. After 24 hours, the cultures were harvested and clarified via centrifugation. The supernatant was collected and the absorbance at 500nm was measured.

3.4.16 - Fed-batch Fermentation.

Fermentations were completed similar to previously reported methods.(B. Pfeifer et al., 2002) Fed-batch aerated fermentations were performed in a 2-liter Biostat A system (Sartorius Stedim SA). An overnight culture was grown in 2 mL of fed-batch medium supplemented with 10ug/mL tetracycline. After reaching late exponential phase at 37°C and shaking at 225 rpm, the culture was centrifuged and resuspended in 50 ml of fed-batch medium supplemented with 10ug/mL tetracycline. The culture was again grown overnight at 30°C and shaking at 225 rpm. 12.5mL of this culture was centrifuged and resuspended in 5 ml of fed-batch medium for inoculation into the Biostat A containing 500mL of fed-batch medium supplemented with 10ug/mL tetracycline. Growth was initially conducted at 37°C with pH maintained at 7.1
throughout the experiment with 1 M H2SO4 and concentrated NH4OH. Agitation was controlled to maintain a level of dissolved oxygen above 30% air saturation.

Glucose concentration was monitored on a 2700 Select Biochemistry Analyzer (YSI). Once the glucose was exhausted from the starting medium, the temperature was reduced to 25°C and IPTG (1mM) and substrate were added. At that point a peristaltic pump started to deliver the feed medium, with glucose maintained below 0.5g/L. Substrate concentration was monitored via HPLC, with another 10mM added approximately every 16 h upon depletion. The fed-batch medium composition (per liter) was: 1.5 g of KH2PO4, 4.34 g of K2HPO4, 0.4 g of (NH4)2SO4, 150.5 mg of MgSO4, 5 g of glucose, 1.25 ml of trace metal solution, and 1.25 ml of vitamin solution. The feed medium composition (per liter) was: 110 g of (NH4)2SO4, 3.9 g of MgSO4, 430 g of glucose, 10 ml of trace metal solution, and 10 ml of vitamin solution. The trace metals solution was composed of (per liter): 27 g of FeCl3 6H2O, 2 g of ZnCl2 4H2O, 2 g of CaCl2 6H2O, 2 g of Na2MoO4 2H2O, 1.9 g of CuSO4 5H2O, 0.5 g of H3BO3, and 100 ml of concentrated HCl. The vitamin solution was composed of (per liter) 0.42 g of riboflavin, 5.4 g of pantothenic acid, 6 g of niacin, 1.4 g of pyridoxine, 0.06 g of biotin, and 0.04 g of folic acid.

3.4.17 - Fermentation Metabolite Extraction and Detection

Liquid-liquid extraction was used to extract metabolites of interest from fermentations. For Ac-Mal<sub>on</sub>-Mal<sub>on</sub>, the fermentation was clarified by centrifugation, and the 500mL supernatant was collected. The supernatant was extracted twice with 500mL HPLC-grade dichloromethane. The organic layers were collected, pooled, and concentrated to 1.5mL by rotary evaporation.
1 µL of the Ac-Mal$_{or}$-Mal$_s$ solvent extract, negative control solvent extract, or d-hexalactone authentic standard (Alfa Aesar) was directly injected into a Zebron™ ZB-5HT, GC Cap Column (30 m x 0.25 mm x 0.25 µm, Phenomenex) in a Trace1310 GC (Thermo Fisher) coupled to an ISQ-LT Mass Spectrometer (Thermo Fisher). The GC inlet was set at 280°C, and split ratio was 10:1. The oven temperature was programmed from 100°C to 250°C at 20°C/min and held for 1 min each at the start and end temperatures. Detection was completed under electron impact mode, with a scan range of 30-350 amu and scan rate 5 scans/second. Transfer line and source temperatures were both 300°C.

3.4.18 - Plasmids

Vector maps and sequences of the Ac-Mal$_{or}$-Mal$_s$, pBR_matABC, and rppA constructs are provided at the following links:

Ac-Mal$_{or}$-Mal$_s$: https://benchling.com/s/seq-oixoOd3X4VLkoEl2RMGM

pBR_matABC: https://benchling.com/s/seq-jhyO1vNmvFt4oYsZha3h

p_rppA: https://benchling.com/s/seq-NqVD6R53GmtwuVhPEXbD

3.4.19 - Software and figure generation

Plots were generated in Python 2.7 using the matplotlib plotting libraries and Excel for Mac and figures were made using Adobe Illustrator CC.

3.5 - References


Chapter 4

Design extensibility and scaling with massively multiplexed genome engineering.

4.1 - Introduction

The goal of this project was to enable combinatorial biosynthesis for the eventual directed evolution of small molecules. As described in the introduction, the success of directed evolution experiments is dependent on the diversity in the initial libraries. In theory, diversity can be generated at each of the three levels of the hierarchical architecture in modular megasynthases: the order of modules within the gene, the function of the modules, and the substrate specificity of the modules. Thus, for our approach to be effective for combinatorial biosynthesis it must be extensible, allowing for the design of functional modular megasynthases from different classes. This means demonstrating that we can design linkers for not just PKSs but also NRPSs and PKS-NRPS hybrids. We must also be able to generate diversity at the last level, by altering chassis substrate specificity, dictated by the A cyltransferase (AT) domain in PKS modules and Adenylation domains (A) in NRPS modules. These considerations require that our linkers are not only context-independent at the module level but also, the enzyme level. In other words, the linkers function regardless of what module they are implemented in and/or what enzyme they are implemented in.
In chapter 3, we used a reverse approach to build a specific PKS to produce d-hexalactone because it had a commercially available authentic standard. However, demonstrating the ability to generate diverse libraries of small molecules demands producing molecules that do not have authentic standards. This means using a forward engineering approach to megasynthase design in which the enzyme is designed, and the molecule produced is predicted from the design. This makes the Test portion of the DBT cycle dramatically more challenging, as without a positive control, it impossible to know whether an extraction method, detection method, or both, are working.

Once the ability to generate diversity at these levels has been demonstrated, the last challenge is to show that the ability to build modular megasynthases scales. No matter how effective the design approach is, the ability to build libraries must be able to compete with traditional methods. This means scalability. As mentioned, diversity can be generated at any of three levels, the order of modules within the gene, the function of said modules, and the substrate specificity of the modules. The structure and order of modules within the gene are introduced during the build step, so assembly must scale. In addition, modification of substrate specificity must also scale.

4.1.1 - This work

To demonstrate the extensibility of this approach to produce a range of other target molecules, we designed and built chimeric Non-Ribosomal Peptide Synthetase (NRPS) and NRPS-PKS chassis. We then developed a metabolomics approach for the identification of the predicted target molecules. Lastly, we showed that precise de novo design of these chimeric megasynthase chassis allows for the application of massively multiplex genome engineering tools
for further rational diversification, providing a scalable, high efficiency approach to combinatorial biosynthesis.

4.2 - Results

4.2.1 - Design of non-ribosomal peptide synthetases (NRPSs) and PKS-NRPS hybrids

After successful demonstration of de novo design and assembly of our chimeric PKS megasynthase chassis, we next wanted to assess the potential for combinatorial biosynthesis by designing other classes of chimeric megasynthases, NRPSs and PKS-NRPS. To achieve this, we designed two new modules: one, an isoleucine-specific loading module and two, a serine-specific extension module. Our complete set of modules (including those from chapter 3) is included in Figure 4.1.

This set of modules could be combined in 18 different ways, generating a design space of 18 different modular megasynthesase chassis and the corresponding metabolites across four classes:
PKSs, NRPSs, and two different hybrid classes, hybrids with one PKS module and two NRPS modules as well as hybrids with two PKS modules and one NRPS module (Figure 4.2).

**Figure 4.2 | Trimodular megasynthase chassis design space accessible from modules in 4.1.** Each node is a molecule that represents its underlying chassis. Each edge represents a single module difference. Two molecules connected by an edge are produced by modular megasynthases with only one different module. Molecules are grouped
by class, PKSs in orange, PKS-NRPS hybrids with two PKS module and one NRPS module in red, PKS-NRPS hybrids with one PKS module and two NRPS modules in blue, and NRPSs in green. Target molecules we designed and tested are in black along the right edge.

We sought to demonstrate that our design approach worked for each of these classes. We chose to do this by traversing through the design space; first, we built the design Ac-Mal$_{on}$-Ser, followed by Ile-Mal$_{on}$-Ser, and finally Ile-Ser-Ser. The structures of the molecules produced by these designs are presented in Figure 4.2.

4.2.2 - Genome Integration, Genome Sequencing, Proteomics, and Fermentation of Ac-Mal$_{on}$-Ser, Ile-Mal$_{on}$-Ser, and Ile-Ser-Ser

Design, cloning, genome integration, genome sequencing, and proteomics were performed as described for the δελτα-hexalactone-producing PKS. The results were similar to those for the δελτα-hexalactone-producing PKS. Each megasynthase was sequence verified and was well expressed upon IPTG-induction, as was sfp. Each new strain was fermented as for δελτα-hexalactone production, with different feed mixtures based on required substrates.

4.2.3 - Computational Chemistry for Metabolomics

As described in section 4.1, unlike δελτα-hexalactone, the target molecules from the new designs presented a more difficult detection challenge because they had no existing authentic standards. Identifying an appropriate detection method requires at least a basic understanding of the properties of the given molecule. For example, from the known boiling point of δελτα-hexalactone, 231°C, we knew it would be compatible with GC-MS. In addition, its lack of polar
functional groups suggested it would be compatible with a highly non-polar solvent for liquid-liquid extraction.

Usually, chemical properties are determined experimentally but, without an available standard, this is impossible. Modern computational techniques however have begun to enable prediction of chemical properties instead of empirical determination. (Goll & Jurs, 1999; Lipinski, Lombardo, Dominy, & Feeney, 2001) Thus, for the predicted metabolites produced by Ac-MalOH-Ser, Ile-MalOH-Ser, and Ile-Ser-Ser, we used available computational tools to predict properties. Specifically, we used iLab from Advanced Chemistry Development Labs. The relevant predictions are presented in Table 4.1.

Table 4.1 | Target molecule chemical properties. Boiling point, logP, and logD properties as calculated by ACDLabs iLab.

<table>
<thead>
<tr>
<th>Design name</th>
<th>Boiling Point (C)</th>
<th>logP</th>
<th>logD (pH 0.1 – 14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ac-MalOH-MalH</td>
<td>231</td>
<td>0.29 ± 0.28</td>
<td>0.29 ± 0.28</td>
</tr>
<tr>
<td>Ac-MalOH-Ser</td>
<td>502.46 ± 35.00</td>
<td>-2.83 ± 0.42</td>
<td>-2.86 – -1.92</td>
</tr>
<tr>
<td>Ile-MalOH-Ser</td>
<td>524.13 ± 35.00</td>
<td>-2.25 ± 0.45</td>
<td>-4.47 – -1.38</td>
</tr>
<tr>
<td>Ile-Ser-Ser</td>
<td>671.93 ± 55.00</td>
<td>-2.74 ± 0.65</td>
<td>-7.55 – -2.55</td>
</tr>
</tbody>
</table>

The predicted boiling point for each molecule was well beyond what is compatible with GC-MS, with each boiling point over 500C. We hypothesized that this was due to the increased functionality compared to δελτα-α-hexalactone causing increased polarity and hydrogen-bonding. As a result, we chose to perform an LC-MS/MS metabolomics approach.

The behavior of molecules during liquid chromatography is dictated by their polarity. The most commonly used metrics for polarity are logP and logD. Both measure partitioning of a molecule between an aqueous phase and a lipophilic phase, such as octanol. The difference
between the two metrics is that logP is for non-ionizable compounds (i.e. does not take pH of the aqueous phase into account) while logD is for ionizable compounds and is thus pH dependent. For each of these metrics, the more polar the molecule, the more negative the value. The predicted values of logP and logD can be used as a starting point for developing chromatographic methods for novel compounds lacking authentic standards. As each of our target molecules was ionizable, we used logD as predicted by iLab from ACD Labs as an estimate of polarity. The plots of logD as a function of pH are presented in figure 4.3. As each molecule had a fairly negative logD, this suggested that the molecules were somewhat polar.

![LogD as a function of pH](image)

**Figure 4.3 | LogD as a function of pH.** Plot of logD as a function of pH for each of our target molecules.

### 4.2.4 - Metabolite Extraction and Detection

From the work detecting d-hexalactone in chapter 3, it was clear that the logD predictions suggested another challenge for detection. For d-hexalactone, our liquid-liquid extraction method
included a significant concentration step (~500X) by rotary evaporation of the non-polar solvent. Without this step, the d-hexalactone concentration would have fallen below the detection limit of our method. Concentration of aqueous solutions is dramatically more difficult, as water cannot be removed via evaporation. In theory, large volume solid phase extraction followed by a lower volume elution is possible but, we did not have access to the required equipment. Instead, we chose to again perform a liquid-liquid extraction but this time with a slightly more polar solvent, ethyl acetate, rather than dichloromethane.

We began LC-MS/MS experiments with reverse phase chromatography, and as each of the compounds is a lactone, we hypothesized they may be present in either a lactone or acid form, depending on equilibrium kinetics of cyclization. (Horsman, Hari, & Boddy, 2015) Using Multiple Reaction Monitoring (MRM), we were able to identify compounds in the Ile-Mal_{OH}-Ser and Ile-Ser-Ser samples with the same masses as the structures in Figure 4.2, and that fragmented into unique product ions (Figure 4.4a and 4.4b) not present in the negative control samples.

![Figure 4.4](image_url)

**Figure 4.4 | LC-MS/MS MRM chromatograms.** A) Ile-Mal_{OH}-Ser fermentation extract (orange) as well as negative control fermentation extract (blue). B) Ile-Ser-Ser fermentation extract (orange) as well as negative control fermentation extract (blue).

The Ile-Mal_{OH}-Ser sample was identified in the lactone channel however, contains two novel peaks with the same MRM transition. This is common behavior for lactones, the acid and lactone forms
separate during chromatography with the acid converting to a lactone in the source of the MS, and has been noted in previous studies of statins (unpublished data). The Ile-Ser-Ser sample was identified in the acid form. This could be due to instability of the lactone form under the selected extraction conditions or, due to an inability of the TE domain to perform the lactonization on this specific substrate.

Unfortunately, we were unable to detect the product of the Ac-MalOH-Ser design. Initially, we hypothesized that the polarity of the molecule may lead it to elute in the void volume of a reverse phase experiment. To test this, we also performed HILIC chromatography and were still unable to detect the target, eliminating chromatography as the problem. The failure to detect the target could be due to numerous factors: first and foremost, the design may be non-functional. Another potential explanation is inefficient extraction as the Ac-MalOH-Ser construct may not have efficiently partitioned into ethyl acetate or it was more volatile than predicted and evaporated during rotary evaporation. Another possibility is that the target does not ionize well in the ESI source of the MS or, does not fragment in the collision cell.

4.2.5 - CREATE Editing

After demonstrating the ability to design and build different functional modules and mixing and matching said modules to build chassis from different modular megasynthase classes, the final step to generate diversity for combinatorial biosynthesis is to alter module substrate specificities. As previously described, substrate specificity of modular megasynthases is dictated by acyltransferase (AT) domains for PKSs and adenylation (A) domains for NRPSs. These catalytic domains are well characterized and have been shown to be amenable to engineered changes in
substrate specificity. (Challis, Ravel, & Townsend, 2000; Dunn & Khosla, 2013; Eppelmann, Stachelhaus, & Marahiel, 2002; Haydock et al., 1995; Rausch, Weber, Kohlbacher, Wohlleben, & Huson, 2005; Reeves et al., 2001; Stachelhaus, Mootz, & Marahiel, 1999) There are two approaches to this engineering. The first is to completely swap catalytic domains and the second is to use targeted mutagenesis of residues implicated in substrate specificity. A major complication with both approaches is the immense diversity, making *in vitro* cloning of every possible module untenable. As an example, by only modifying substrate specificity a single chassis (i.e. a node in figure 4.2) could be diversified to over 20 new chassis (figure 4.5).

![Figure 4.5](image_url)  
*Figure 4.5 | Example of small molecule design space accessible by altering substrate specificity of a single node (or chassis).*

We were able to apply a precise and scalable genome engineering tool developed in our lab, CRISPR-enabled trackable genome engineering (CREATE), to alter chassis substrate specificities in a scalable manner (Figure 4.6). (Garst et al., 2016)
Figure 4.6 | CREATE method at the genome-level and protein-level. A CREATE plasmid contains a CRISPR-RNA and an editing cassette in series. In this case, the CRISPR-RNA targets the serine-specific Adenylation domain in our design Ac-Mal_OH-Ser for cleavage. The editing cassette is a glutamate-specific Adenylation domain that, via homology-directed repair, repairs the cleaved DNA, replacing the serine-specificity for glutamate specificity. After applying CREATE, the serine-specific Adenylation domain (orange) of the design Ac-Mal_OH-Ser is converted to a glutamate-specific Adenylation domain (gray), yielding a new design (Ac-Mal_OH-Glu). This design produces a new molecule.

Briefly, CREATE is a CRISPR-based technology that involves synthesizing constructs which contain an editing cassette and CRISPR-RNA sequentially. Previously, these constructs have been synthesized in massive multiplex on an oligonucleotide array. We tried this technology for both approaches to engineering substrate specificities.

First, we wanted to modify the substrate specificity of the adenylation domain using the fewest possible mutations. This meant using the Stachelhaus code as a guide to design a minimal cassette. (Stachelhaus et al., 1999) Using the Stachelhaus code, we chose to swap serine specificity for glutamate specificity, as this only required five amino acid changes. As part of the minimal cassette to make these swaps, we tested four different gRNAs that targeted the adenylation domain. The efficiency of cutting is affected by the gRNA, and the cutting efficiency has been implicated in editing efficiency (data not shown). Unfortunately, none of the cassettes were able to successfully edit the site. Sequencing data showed that when editing did occur, there were significant crossover events between partial regions of the cassette and the genome, giving
incomplete editing. We hypothesize that this occurred because there was significant homology between the cassette and the genome in regions between mutations.

We tested complete catalytic domain swaps in parallel. Historically, identification of exact splice points at which to swap AT or A domains while maintaining function has been a challenge. (Kim, Moore, & Yoon, 2015; Winn, Fyans, Zhuo, & Micklefield, 2016) However, with a bottom-up or de novo designed chassis, the splice points to make AT or A domain swaps are known. We hypothesized that this fact combined with the precision of the CREATE method would allow us to make constructive or functional swaps. To test this hypothesis, we synthesized a CREATE-style cassette with a CRISPR-RNA that targets the serine-specific adenylation domain of our NRPS extension module and a glutamate-specific adenylation domain as the editing cassette (Figure 4B and C).

As with the minimal cassette, to maximize editing efficiency, we began by testing the same four gRNAs. The sequence and editing efficiency for each of these gRNAs is presented in Table __. Using the best cassette, we swapped serine specificity for glutamate specificity (Ac-MalOH-Ser to Ac-MalOH-Glu) in our design. The sequence verification of this swap is presented in Figure 4.7.
Figure 4.7 | Sanger sequencing results. Successful swap of Glutamate-specific Adenylation Domain (gray) for Serine-specific Adenylation Domain (orange) using CREATE-style approach. The linkers (red) on the 5’ and 3’ ends of the adenylation domain are unchanged.

This design was tested as described for the other designs and shown to produce a compound at the expected mass of the lactone form. This compound again had two peaks, as described for the Ile-Mal$_{\text{OH}}$-Ser sample (Figure 4.8a). Initially, there appeared to be tiny peaks in the negative control at these retention times (Figure 4.8a). To test whether these peaks were signal or noise, we performed the same LC-MS/MS experiment on a more sensitive instrument (an ABI SCIEX 6500 vs 4000). With the increased sensitivity, signals should rise while noise remains the same (i.e. increased signal to noise ratio). We hypothesized that if the peaks present in the negative control were signal, the ratio of the negative control peaks to the the peaks in the experimental samples would remain the same (i.e. both signals would rise with the boost in sensitivity). Instead, we saw that the small peaks in the negative control were unchanged, while the signals in the experimental samples increased dramatically (Figure 4.8b). This confirms that the peaks in the experimental sample were signal, while peaks in the negative control are noise.
4.2.6 - Library scale.

After demonstrating that our approach is broadly applicable to designing functional modular megasynthases and that it enables generation of diversity at all levels, the next step is to begin assembling libraries.

4.2.6.1 - Multiplexed chassis assembly

To begin the process of assembling a library, we decided to build all of the modular megasynthase chassis accessible from our parts. In other words, we mixed and matched all of our designed modules from figure 4.1 to generate the 18 chassis in figure 4.2. As we placed an emphasis on scalability, we set out to clone each of these chassis in a single experiment.

To do this, we used PCR to install all possible homology arms on each module. This generated pieces that allowed for installation of each module in every position within the trimodular megasynthase. We pooled these pieces for a single multiplexed yeast assembly. The
library was successfully assembled. Ideally, these constructs could remain pooled throughout the entire assembly process, all the way to genome integration. Unfortunately however, transforming the assembled constructs into *E. coli* cloning strains as described in chapter 3 collapses library coverage. To circumvent this problem, we demultiplexed the library after yeast assembly, by picking colonies and identifying assemblies with junction PCRs (Figure 4.9). Each construct was then moved through the rest of the process, eventually being integrated into the *E. coli* genome.

![Figure 4.9 | Example junction PCR for demultiplexing. Highlighted in red are colonies with positive results for both expected junctions. Most colonies have one of the junctions but, not both.](image)

4.2.6.2 - CREATE Library

In parallel, we prepared a CREATE library to modify substrate specificities of this library. To do this, we made CREATE constructs to swap the serine adenylation domain with adenylation domains specific to the 19 other amino acids as described in section 4.2.5. In one CREATE experiment, this set of constructs alone would expand our library from 18 chassis to 284 modular megasynthases. This library did not include cassettes for changing substrate specificity of acyltransferases, as *E. coli* only produces acetyl and malonyl-CoA derivatives. Any other acyl-CoA derivatives would require further metabolic engineering of the background strain for substrate feeding.
4.3 - Discussion

In this chapter, we have demonstrated that the modular megasynthase design approach outlined in chapters 2 and 3 is extensible to modular megasynthase classes beyond Polyketide Synthases (PKS), specifically Non-ribosomal peptide synthases (NRPSs) and PKS-NRPS hybrids. This suggests our linker designs are also context independent at the enzyme-level. Again, as an example, the KS-AT linkers used across each chassis presented here are the same and maintained function.

In addition, we have shown that our designs were compatible with a scalable genome engineering tool, allowing us to make a constructive adenylation domain swap to alter the substrate specificity of a module within a megasynthase. We hypothesize that the precision of our designs gave us the prior knowledge to know exactly where to splice when making this swap, and that it is the precision of the swap that it enabled it to be functional.

Combined, these results have shown that our design approach is able to produce modular megasynthase designs that vary at all three levels, the structure of modules, the order of modules, and the substrate specificity of modules. This makes our design approach an ideal method for combinatorial biosynthesis. We tested our ability to implement this approach at scale, as library size is directly implicated in the success of directed evolution experiments.

We were able to successfully assemble a small library of chassis in a multiplexed yeast experiment however, the intermediate steps between yeast assembly and genome integration proved to not scale. For future experiments, this would have to be resolved by further optimizing
the *E. coli* transformation process. We also showed that once the chassis library was built, it could be further diversified at scale using CREATE.

The scalability of CREATE could be improved. The first implementations of CREATE involved synthesizing cassette libraries on an oligonucleotide array. This is a very cheap and scalable way of building libraries but the cassettes on an oligonucleotide array are maximum 250bp in length. The cassettes we used to swap catalytic domains were ~2.2kb in length, and thus are not compatible with multiplexed synthesis.

Targeting the specific residues implicated in substrate specificity would shrink the size of the cassettes required, assuming the residues are not distributed across the entire catalytic domain. We did attempt to use smaller cassettes that only mutated specific residues but, these failed. Another way to target specific residues would be to iteratively saturate each residue one at a time. This would be compatible with multiplexed array-based oligonucleotide synthesis however, the editing efficiencies of CREATE are much too low. As an example, our library should generate 284 megasynthases. If this were assembled by iteratively saturating 10 sites at 12.5% editing efficiency, after 10 rounds only one in a billion cells would have all sites completely saturated ($0.125^{10} = 9.3 \times 10^{-10}$).

Finally, although we demonstrated the ability to build large libraries of these constructs, we did not test this library for function. Test was a significant limitation here for two reasons. For one, performing 284 100-hr fermentations is not feasible. A second option would be to perform one multiplexed fermentation of all 284 strains, but this would be immensely difficult to test, as the concentration of each metabolite would be so low, they would likely be near or below their detection limit. In addition, without authentic standards it would be extremely difficult to design a metabolomics approach that would cover 284 molecules, each with different chemical properties.
4.4 - Materials and Methods

4.4.1 - Design, Build, and Test

Performed as described in chapter 3.

4.4.2 - CREATE-Style Genome Editing

Our CREATE-style genome editing cassette contained a 500bp homology arm, the Glutamate-specific adenylation domain, a 500bp homology arm, the 35nt constitutive promoter J23119, a 20nt spacer, and 24nt of the 5' side of the standard S. pyogenes gRNA. This cassette was synthesized as a gBlock by IDT and cloned into the gRNA backbone reported by (Qi et al., 2013) and then used to edit the Serine-specific adenylation domain in our Ac-Mal\textsubscript{OH}-Ser chassis. The protocol for editing was as previously described. (Garst et al., 2016)

4.4.3 - Plasmids

Vector maps and sequences of the Ac-Mal\textsubscript{OH}-Ser, Ile-Mal\textsubscript{OH}-Ser, and Ile-Ser-Ser constructs as well as the Ser/Glu\_CREATE plasmid are provided at the following links:

Ac-Mal\textsubscript{OH}-Ser: https://benchling.com/s/seq-GLNizgd9QzQcspSLx7j
Ile-Mal\textsubscript{OH}-Ser: https://benchling.com/s/seq-heSYqew01abNw4PWhUd
Ile-Ser-Ser: https://benchling.com/s/seq-i6p1LGWgCm3rpzeS144l
4.4.4 - Metabolite Extraction

For Ac-Mal$_{\text{ser}}$-Ser, Ile-Mal$_{\text{ser}}$-Ser, and Ile-Ser-Ser, the fermentations were lysed using a Branson Sonicator at 40% amplitude. The lysate was clarified via centrifugation and the 500mL supernatant collected. The supernatant was then extracted twice with 500mL HPLC-grade Ethyl Acetate. The organic layers were collected, pooled, and washed with 500mL HPLC-grade water. The washed organic layer was collected and removed by rotary evaporation. The extract was resuspended in 15mL 9:1 HPLC-grade water:HPLC-grade acetonitrile.

4.4.5 - LC-MS/MS Metabolite Detection

Initially, the solvent extracts for Ac-Mal$_{\text{ser}}$-Ser, Ile-Mal$_{\text{ser}}$-Ser, and Ile-Ser-Ser were directly infused into a AB Sciex API 4000 Tandem Mass Spectrometer and Product Ion mass spectra of the samples and negative control were recorded using the positive product ion scan mode of the anticipated parent ion masses [M+H]$^+$ at a collision energy (CE) of 15 – 25eV and declustering potential (DP) of 25V. Product ions unique to the test samples were selected to generate multiple reaction monitoring (MRM) transitions. For LC-MS/MS analysis, 5µL of the samples and negative controls were injected on an XBridge™ C18 reverse phase columns (100 mm x 2.1 mm x 3.5µm, Waters). The mobile phases of Acetonitrile and LC-MS grade water with 0.02% formic acid (aqueous=A) were run at 0.600mL/min with a gradient starting at 95% A for one minute and ramped down to 2% A within 3 min.
4.4.6 - Multiplexed Yeast Assembly

Multiplexed yeast assembly was performed as described in chapter 3 however, all modular pieces were pooled for a single experiment. The plated transformation was demultiplexed by selecting colonies and genotyping them with junction PCR to identify which assembly was present.

4.4.7 - CREATE Library

To build a CREATE library of adenylation domains, we used the same genes reported in Stachelhaus et al. (Stachelhaus et al., 1999) The adenylation domains were annotated using antiSMASH 2.0. (Blin et al., 2013) Each adenylation domain was codon harmonized to a CAI of 0.7 and then synthesized by IDT with 50bp homology arms to the Ser/Glu CREATE backbone (section 4.4.2). The Ser/Glu CREATE backbone was linearized with PCR, and the cassettes were cloned using the Gibson Assembly HiFi 1-Step Kit (Synthetic Genomics International, Inc.). The plasmids were sequence verified. Finally, the CREATE experiment was performed as described in section 4.4.2.

4.5 - References


Chapter 5

Conclusions and Future Work

5.1 - Conclusions

We have demonstrated the ability to successfully design, assemble, and express functional chimeric megasynthase enzymes from the bottom-up as well as editing of such enzymes in a precise and scalable manner that can be performed massively in multiplex to potentially produce 10,000’s of natural product molecules. This was enabled by a new approach to megasynthase modularity that treats catalytic domain linkers as modular parts, as well as the catalytic domains themselves.

We reported a computational design tool that yields putative catalytic domain linkers. The sequences output by this tool are overrepresented in each sequence database. We speculate that these sequences are conserved and may have the desired context-independence because they are functional but have enough structural flexibility to be context-independent. This result is consistent with the hypothesis that as structural plasticity decreases, so does promiscuity and evolvability.

These linkers, when combined with catalytic domains, enabled de novo design of multiple modules and complete chimeric modular megasynthase chassis, suggesting that the linkers are context-independent. At the module level, the same linker amino acid sequence was used in multiple modules; for example, the KS-AT linker amino acid sequence was the same in both extension modules and both modules functioned as intended. The linkers also appear to be context
independent at the enzyme-level. Again, as an example, the KS-AT linkers used across each chassis presented here are the same and maintained function.

This approach appears to have broad scope and high-efficiency, as it was successfully applied to three classes of modular megasynthase: PKSs, NRPSs, and PKS-NRPS hybrids. Historically, implementation at scale has been mostly limited by molecular biology methods that are either precise but not scalable or, scalable but not precise. We have shown that our chassis design method is compatible with a novel, massively multiplexed genome engineering tool that is both precise and scalable. Together, these results suggest that our design approach can enable facile assembly of large, diverse libraries of functional modular megasynthases.

5.2 - Future Work

Future work should include building modular megasynthases of larger size than the minimal trimodules demonstrated here, adding new parts such as tailoring domains like O- and N-methyltransferases, and ideally, applying CREATE to build libraries orders of magnitude larger than previously reported.

As our ability to design biologically synthesized, novel, complex molecules expands, library quality control will become increasingly important. Our results demonstrate the significant challenge that will come with large, diverse small molecule libraries: a lack of authentic standards and varied compatibility with methods for metabolite extraction and detection. For example, our inability to detect the product of the Ac-Mal<sub>ac</sub>-Ser design could be due to numerous factors: non-functional design, inefficient extraction, poor chromatography, poor ionization, poor fragmentation, etc. As no universal, unbiased, and definitive metabolomics method exists,
improvements in the design and build process as reported here shift the rate limiting step in combinatorial biosynthesis to metabolomics.

The metabolomics challenge can also be viewed from a different perspective, that of Design-Build-Test (DBT) prototyping described in chapter 3 and used in chapter 4. Ideally any design or engineering process is iterative in nature, yielding cumulative small improvements. This concept maps well to the initial thrust of this dissertation, to enable directed evolution of small molecules. In mapping directed evolution onto DBT, design and build steps represent generation of the diverse population and test represents selection of the most ‘fit’ members. The rate at which improvements in design or fitness accumulate is limited to the slowest step in the process. Design can be performed effectively ad infinitum computationally, and DNA synthesis and assembly are improving rapidly. Therefore, for prototyping modular megasynthases, the rate-limiting step is test, or metabolomics. For directed evolution, with massive, multiplexed combinatorial biosynthesis libraries in hand, selection is the limiting factor.

We are left with a dilemma, do we invest resources in building large, multiplexed combinatorial biosynthetic libraries without adequate methods for validation or screening? Or, do we wait to scale until we have test methods that work at equivalent scale?
Bibliography


