Towards sustainable biomolecule production: computational approaches to accelerate genetic tool development for engineering metabolism in microorganisms

Erin H. Wilson

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

David Beck, Chair

Mary Lidstrom

Larry Ruzzo

Program Authorized to Offer Degree:

Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Towards sustainable biomolecule production: computational approaches to accelerate genetic tool development for engineering metabolism in microorganisms

Erin H. Wilson

Chair of the Supervisory Committee:
David A. C. Beck
Department of Chemical Engineering

Globally, human societies are consuming finite resources at unsustainable rates. Transitioning away from our dependencies on non-renewable resources and towards cyclical production of everyday materials is critical for mitigating our escalating impact on climate change and securing longer term economic stability. A promising alternative to sourcing many materials is via metabolic engineering: a field that aims to engineer microorganisms into biological factories that convert renewable feedstocks into valuable molecules (i.e., jet fuel, medicine, bioplastics). In order for metabolic engineering solutions to be economically viable, microorganism factories must be optimized to produce target molecules quickly and at high yields. Such optimization requires an understanding of the complex genetic grammar that controls gene expression within a host microbe as well as genetic tools with which to manipulate it. While extensive genetic toolkits have been developed for model systems like *S. cerevisiae* and *E. coli*, many non-model organisms lack tools with which to effectively engineer them.

This dissertation explores computational approaches for developing genetic tools in non-model microbes, using the methanotroph *Methylotuvimicrobium buryatense* as an example. First, we discuss a framework that leverages RNA-seq datasets to predict constitutive, strong promoters, which we developed into a suite of expression tools in *M. buryatense*. Next, we use unsupervised machine learning methods to identify 43 independently modulated groups of co-expressed genes (iModulons); interactively explorable visualizations of these data facilitated a deeper characterization of *M. buryatense* expression modules across diverse growth conditions and a proposed set of gene candidates for functional validation via mutation experiments. Finally, we investigate the potential of deep learning models to predict gene expression behavior directly from *M. buryatense* promoter sequence regions and probe the performance limits of common model architectures in varied genomic contexts and data-limited regimes.

This work contributes to a broader understanding of how computational techniques can be used to model the effects of biological sequences on gene expression outcomes and describes scenarios in which these techniques are more limited. Such guidance will enhance our collective ability to use computational approaches for genetic tool development in non-model microbes and accelerate metabolic engineering solutions that can shift humanity towards a more sustainable relationship with our planet's finite resources.

Table of Contents

Acknowledgements

Grad school is a journey, and I'm extremely grateful for so many people who helped me along the way.

To my committee: thank you all for your constructive ideas and encouragement throughout my dissertation. **Jennifer**, thank you for the supportive check-ins leading up to my exams and for your thoughtful feedback. I was so glad to have your additional synbio expertise on my committee! **Sara**, thank you for letting me moonlight in your lab meetings and helping me find a closer home within the CSE CompBio community. The extra ideas from you and your lab really helped energize me to get to the end! **Larry**, thank you for always making time for me whenever I reached out and for listening with care, whether I was sharing some exciting news or some worries. Your words of wisdom have helped me tremendously throughout grad school! **Mary**, I have loved having you as my personal microbio safari guide into the world of methanotrophs! Thanks for taking a chance on a computational grad student, for always engaging with my ideas, and for your generally warm guidance as I navigated the ups and downs of research. **Dave**, thank you for always nerding-out with me, whether it is about data science, open-source software, or dungeons and dragons! Your enthusiasm for science is contagious, your wizardry on the command line is inspiring, and I always felt more energized to try new ideas after talking with you. To both my advisors, **Dave and Mary**: someone once told me, "Grad school is like Jedi training: you have to find your Yoda." Thanks for being great Yodas, without demanding piggy back rides through swamps :) I've learned so much from you both!

To the **Lidstrom, Beck, Mostafavi labs**: thanks for all your help brainstorming and troubleshooting - science was much more fun with you all around! And thanks to **Lars** for hosting me in Denmark, and to everyone in the **Nielsen lab** for welcoming me at DTU Biosustain and including me in your riveting *kantine* lunch discussions!

To the **CompBio peeps,** the **#paul-gs-ogs cohort,** and the whole **UW CSE community**: you all have been such a goofy crew to share grad school with - thanks for all the laughs, the hikes, the picnics, and for tolerating my many Star Wars-themed events. What a wonderful and vibrant support network! An especially big hug to the "2017 CompBio Ladies," **Alex, Ayse, Lee**, and **Nicasia**: thanks for helping me survive and thrive - y'all are truly the best. I can't wait to see where your stories go from here <3

So many friends and family beyond UW have been essential to keeping me smiling and supported along this journey, whether we've known each other since we were kids or only met this past year! From teammates on the **Quakes/Mavericks/Wildcats/Knights** who's desire, dedication, and discipline continues to inspire me to find more within myself; **Carleton nerdfolk** who keep the magic of Northfield alive wherever we find each other; **family** who send me fun notes of encouragement from afar; **Frizbee friends** who help me find my chill; **Seattle**

housemates who share meals and help me feel at home in the PNW; **Perry and the porgs**, who's cute and charismatic nature captivated a larger audience than my own research ever did; and **D&D companions** who spark my imagination and have my back whenever we venture into fantastic unknowns

To **Mom, Dad,** and **Bryn**: all your gestures of support, big and small, have been so appreciated these past 6 years. Thanks for believing in me and sending me cookies whenever I needed them most :) Thanks to **Charlez** for the much needed puppy snugglez. Thank you to the **Johnsons** for welcoming me into your family and cheering me on alongside my own family.

To **Matt**: you rock. Your stealthy snack deliveries and curious questions kept me fueled and thinking more deeply about my research. Thank you for striking the perfect balance of supporting me in my PhD journey while giving me the space to become my own scientist. I would like to spend more time elaborating on my admiration for you as a compassionate human who builds technology to connect people and ideas, but I guess I should wrap this up and turn it in so we can go plan a wedding…

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Synthetic biology for sustainability

As the world's population continues to increase, so will its resource requirements. To better manage these resources for generations to come, our use of energy, land, and materials must become more cyclical: our inputs drawn from renewable sources and our outputs destined for more than an accumulation of waste.

Synthetic biology is a growing field that offers a promising alternative to sourcing many materials more sustainably.[1,2] Sitting at the intersection of scientific disciplines such as biology, chemistry, computer science, math, and more, synthetic biology aims to engineer novel biological systems to solve pressing global challenges in areas such as medicine, conservation, and sustainability.[3] While the applications of synthetic biology are quite broad, this dissertation will focus on one particular goal: *sustainable biomolecule production.*

Humans rely on many biologically-derived molecules: fuels for transportation, fibers in clothing, medicinal molecules from plants. Molecules found in nature are typically produced by organisms that can execute a specific metabolic pathway, or a series of chemical conversions carried out by enzymes that can transform inputs the organism consumes into other building blocks it needs to survive. Organisms store instructions for these metabolic pathway enzymes in their genomes as DNA. Since DNA is a common language between all organisms, genetic instructions are potentially transferable between species.

Metabolic engineering is a subfield of synthetic biology that leverages this genetic transferability of enzyme instructions to rewire chemical conversion pathways in microorganisms, like bacteria, to produce a range of valuable molecules that other organisms, like plants, make naturally.[4,5] By installing heterologous enzymatic pathways into a microbial host and optimizing the host metabolism to funnel its input resources down these pathways, we can create synthetic, biological factories for molecules of interest (Figure 1.1).

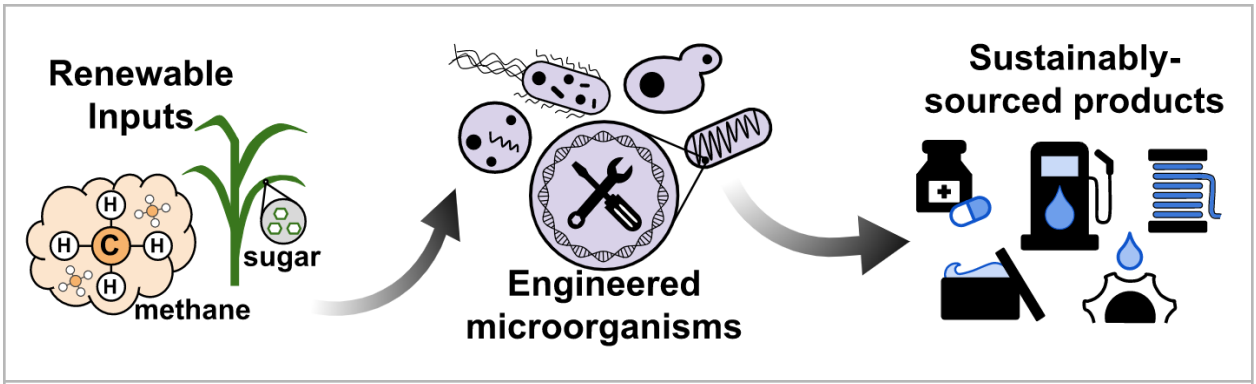


Figure 1.1. Overview of metabolic engineering: a process to engineer microorganisms into biological factories that convert renewable inputs into more sustainably-sourced products.

Engineering biological factories that can produce valuable molecules efficiently and at large scales is a promising avenue for sourcing essential materials more sustainably and with reduced emissions.[6,7] Especially if the inputs to these biological factories are renewable resources (i.e., sugar cane) or a waste stream from some other manufacturing process (i.e., greenhouse gas emissions, steel mill off-gas), it provides an alternative to sourcing materials typically derived from fossil fuels or other ecologically destructive processes. Not only could such innovations contribute to increased economic stability by reducing reliance on delicate geopolitical relationships for oil, but additionally it could improve environmental health outcomes by reducing the amount of fossil-based carbon being introduced into the global carbon cycle or by diverting emissions streams out of the atmosphere.[8,9] However, input sources such as sugar cane come with additional complexities: though renewable, there exists competition with land for food production as well as land for natural ecosystems. Such tradeoffs must be considered when assessing the overall impact of synthetic biology solutions for sustainability.[10]

Excitingly, there have already been many metabolic engineering success stories. Notable efforts include the use of *Saccharomyces cerevisiae* to convert sugar cane into farnesene (jet fuel)[11,12] and artemisinic acid (anti-malaria drug precursor)[13,14], the use of *Escherichia coli* to produce the polymer building blocks 1,4-butanediol[15] and 1,3-propanediol[16], a *Salmonella*-based control system for secreting spider-silk,[17] and engineered *Pichia pastoris* that produces soy leghemoglobin, a meaty flavor added to plant-based meat alternatives.[18] In recent decades, an entire ecosystem of synthetic biology companies has emerged as the field continues to drive progress towards bio-based production and commercialization of molecules we use in everyday life.[7]

Microbe optimization remains challenging, however. One major hurdle is that it requires finely-tuned expression levels of each heterologous gene, but we do not yet fully

understand the rules that govern gene expression. Efforts such as altering protein codon frequencies to match the host microbe's preference is critical for achieving successful expression[19]; however, genes are further regulated by an ecosystem of non-coding genetic signals such as promoter and terminator elements, 5' and 3' untranslated regions, enhancers, and more.[20,21] A deeper understanding of the "genetic grammar" that underlies the regulatory signaling patterns in an organism is critical for engineering new pathways into microbes with more predictable expression.

A major mechanism through which microbes control gene expression is transcription initiation, a process largely influenced by sigma and transcription factors binding to regulatory DNA elements in promoter regions.[22–24] Though organisms have evolved multiple layers of regulatory mechanisms such as mRNA degradation, post-translational modifications, and protein-protein interactions,[20] transcription initiation is a critical first step and will be the primary focus of this dissertation.

1.2. Developing genetic tools for non-model organisms

Microbial hosts such as *S. cerevisiae* and *E. coli* serve as excellent platforms for metabolic engineering: their roles as model organisms have led to detailed characterizations of their genetics and physiology, and many genetic tools have already been developed for engineering them.[5,6] With proper tools that enable scientists to control and reprogram the gene expression, organisms may be more effectively optimized to produce target molecules at high titers, rates, and yields. Such optimizations can reduce the overall cost of molecule production and improve the economics enough to compete with platforms rooted in fossil fuels and greenhouse gas emissions.

Though model organisms with established genetic toolkits are relatively easy to work with, there exists a vast pool of alternative microorganisms that may offer significant benefits to molecule production processes.[25,26] The ability to leverage the diversity across microbial life opens a much broader solution space for metabolic engineering approaches to succeed, and ultimately outcompete unsustainable production processes.

Unfortunately, for many non-model organisms, extensive genetic toolkits are not yet available. In some cases, tools developed in a model organism can indeed be transferred to a new organism and remain functional,[27] however it is not guaranteed that parts will be compatible across species.[28] Organisms have evolved intricate systems of controls to regulate gene expression: genetic signals encoded as DNA sequence patterns exist throughout the genome and are often short and can be arranged in many

different combinations and orientations.[24,29,30] While our knowledge about microbial genomes and their regulatory codes has grown significantly with the advent of DNA sequencing technologies, there still exist significant start-up costs to developing suites of genetic tools and thus adopting new organisms as metabolic engineering platforms can be slow.[6]

When working in a new non-model microbial host, metabolic engineers often strive to establish a proof of concept that engineering is feasible in that organism. Especially for organisms for which there are not yet reliable databases available, computational approaches that rely on relatively accessible, simple-to-collect data types rather than requiring highly specialized experiments would reduce the time and investment required to explore their viability as biomolecule production hosts. Strategies that can quickly establish genetic tools, such as promoter constructs, are essential for enabling rapid prototyping of engineered pathways, which in turn expands our ability to explore microbial solution search spaces to meet varied climate change mitigation goals.

1.3. A case for methanotrophic hosts

A compelling set of candidates for metabolic engineering hosts are microbes that can survive on one-carbon (C1) compounds. Many C1 compounds - including carbon dioxide (CO_2), carbon monoxide (CO), and methane (CH_4) - are prominent greenhouse gasses, dangerous pollutants, or even byproducts of industrial waste streams.[31] Microorganisms that naturally consume these molecules play key ecological roles in cycling carbon back into the environment.[32–34] If such organisms could be reliably engineered to produce molecules of interest, the benefits would be two-fold: not only would these molecules be produced via more sustainable biological means, but their polluting feedstocks would be diverted out of the environment and sequestered into valuable materials.[35–37]

Methane is a greenhouse gas that is heavily emitted through both natural sources (wetlands, wildfires, permafrost) and human sources (oil and gas production, cattle ranching, wastewater treatment, rice paddy agriculture, landfills, coal mines).[38] According to the global methane budget distributed by the Global Carbon Project, the rate of global methane emissions exceeds the rate of methane sinks by about 10 million tons of CH_4 per year and as a result, is increasing in the atmosphere.[31] Though less abundant than carbon dioxide, methane is 85x more potent in its warming potential over a 20 year time frame, and thus addressing methane emissions is a critical avenue for mitigating climate impacts.[39–41]

One promising microbe for methane mitigation with synthetic biology is the methanotroph *Methylotuvimicrobium buryatense* 5GB1, a bacterium that can consume methane as its sole carbon source.[32,33] Previous work in *M. buryatense* has examined physiological responses in stress conditions, developed genetic manipulation tools, built genome-scale metabolic models, and measured flux via isotope labeling techniques.[27,42–45] In a screen of seven phylogenetically diverse methanotrophs, *M. buryatense* showed an increased ability to assimilate methane at 500 - 1000 ppm and maintain robust growth (Lian He, unpublished). This concentration of methane is higher than in the general atmosphere (~1.9 ppm[46]), but is comparable to methane concentrations observed in the air above major emissions sites, such as landfills, rice paddies, and oil and gas production sites.[47–49] This presents an intriguing opportunity for methane air-capture technology paired with methanotrophs cultivated in portable bioreactors: if deployed directly at emissions sites where methane concentrations are elevated, emissions could be mitigated at their sources.[50] Furthermore, if engineered methanotrophs could be incorporated into such a system, it could simultaneously capture methane emissions and convert the carbon into valuable products. While many pieces of this technology are still under early investigation and development, in order for such a platform to be economically effective as a mechanism for biomolecule production, a deeper characterization of *M. buryatense's* genetic grammar is required.

1.4. Dissertation overview

With such promising technological innovations on the horizon, this dissertation describes progress towards enabling *M. buryatense* as an engineerable biomolecule production host with a focus on computational methods that are generalizable to other non-model microbes. The following chapters each implement and characterize a computational approach for decoding genetic grammars and discuss the applicability of each method to non-model microbes in data-limited regimes.

In Chapter 2, we describe the dataset used throughout this dissertation: a compendium of RNA-seq samples collected in *M. buryatense* over the last decade in the Lidstrom Lab at the University of Washington. We provide details on the diverse growth conditions under which the data were collected, highlight subsets of particular relevance to synthetic biology, and comment on several previously unknown expression trends observed among non-coding RNAs. Additionally, we discuss its strengths and limitations as a representative dataset for studying non-model microbes.

In Chapter 3, we outline a computational framework for identifying strong, constitutive promoters in non-model organisms. When applied to the *M. buryatense* RNA-seq compendium, we were able to identify a set of top genes which maintained strong

expression across a wide variety of growth conditions, distill a core sigma-70 signal from a set of promoter predictions, and experimentally validate expression for a set of minimal promoter predictions, including a synthetic sequence not found in the native genome. This work resulted in an expanded promoter toolkit for *M. buryatense* and an open-source software tool with interactive visualizations and tutorials demonstrating its application to several microbes.

In Chapter 4, we describe a deeper investigation of co-regulated gene modules in *M. buryatense* using Independent Component Analysis (ICA) - an unsupervised machine learning technique - on the RNA-seq data. This work recapitulated several known expression modules, such as those involved in transcriptional responses to the presence and absence of copper and lanthanum, and provided interactive visualizations in which to explore expression trends for additional groupings of genes that seemingly respond in other types of physiological stress conditions. In the hands of expert microbiologists, this tool led to several proposals for mutation candidates to experimentally validate for gene function.

In Chapter 5, we set out in search of *M. buryatense* regulatory motifs using recently developed deep learning methods for DNA sequence inputs. Though the initial goal was to identify influential motifs that could be developed into genetic tools, such as inducible and repressible promoters, this effort evolved into a deeper characterization of deep learning methods performance in data-limited regimes. We explore the effectiveness of convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) for various prediction tasks using DNA sequences as inputs including 1) predicting RNA-seq expression outcomes from all native *M. buryatense* promoter regions, 2) characterizing the limits of model performance for predicting expression outcomes measured in Massively Parallel Reporter Assays (MPRA) after systematic reductions in training data, and 3) assessing the interplay between signal complexity and dataset size in a synthetic motif repression/activation scenario. From these computational experiments, we observe that the deep learning methods we tested in data limited-regimes are insufficient for learning genetic grammars. Despite attempting several model pre-training strategies for alleviating data limitation, we suggest that additional data beyond that available in microbial RNA-seq is necessary to more fully learn the nuances of microbial genetic grammars. For future researchers pursuing similar motif elucidation queries, we therefore recommend collecting additional data (for example through an MPRA if possible) or pursuing alternative modeling approaches to those described here.

Finally, Chapter 6 concludes with a summary of the efforts described here to accelerate genetic tool development for non-model microbes and key takeaways from using deep

learning in data limited regimes. Additionally, I offer some reflections on the role of communication and accessibility in science, from creating interactively explorable visualizations and tutorials to communicating scientific ideas with general audience-friendly prose. Every project heavily featured interactive visualizations to enhance understanding of the data and enable independent discovery for other users of my tools. Even for projects in which the outcomes were primarily negative results, I was excited to publish tutorials to help others learn. A major goal throughout the work in this dissertation has been to expand the accessibility of ideas, both to people with expertises in adjacent scientific domains and to people outside the realm of science or academia.

Overall, this dissertation captures my effort to bridge across the technical fields of synthetic biology, machine learning, and data science, and I hope it additionally conveys my earnest desire to help shift humans towards a more sustainable relationship with the planet.

Chapter 2. Compilation of *M. buryatense* RNA-seq compendium

2.1. Overview of transcriptomics

Organisms dynamically respond to changing environments by selectively upregulating and downregulating sets of genes to initiate a physiological response. Often this will involve the production of proteins to cope with the changing environment, such as producing enzymes to break down a newly detected energy source or preparing to move towards or away from a stimulus. These response cascades start with transcription: an initiation of cellular machinery to copy gene coding sequences into RNA that is later translated into an amino acid sequence, and eventually folded into a functional protein.[51]

Transcription itself does not guarantee protein synthesis, as there are several other regulatory mechanisms which can inhibit gene expression. However transcription is a key first step and analyzing gene expression data has been informative for identifying groups of genes that participate in coordinated cellular processes.[52] Often this coordination relies on a shared signaling mechanism, such as a binding event of a transcription factor to a particular DNA motif. The promoter region immediately upstream of gene coding sequences is a genomic region where many such motifs are concentrated.[29] Once identified and characterized, signaling motifs can be repurposed for metabolic engineering objectives by incorporating signals into synthetic expression constructs. For example, by pairing native yeast upregulation promoter signals with an artemisinin-production pathway gene from the sweet wormwood plant, we can express the plant gene in the yeast cell's genetic context by recruiting the yeast transcriptional machinery to transcribe the heterologous gene sequence.[13]

Decoding the grammar of gene regulation can be facilitated by RNA-sequencing: a method to measure the current state of mRNA transcript levels in cells. RNA-sequencing provides a temporal snapshot of which genes are being activated or repressed in the current growth conditions, which can be a useful indicator of which metabolic pathways are relevant for a given organism in a given condition and lead to physiological insights.[53,54] From a genetic signaling perspective, RNA-seq measurements can also indicate which sets of genes are being regulated in tandem and therefore identify sets of promoters in which to look for shared motif signals. However, disentangling these signals is complicated by the fact that some genes are expressed together in specific growth conditions, and not in others. It is therefore possible that a specific expression signal is shared between two genes, however if the transcriptional

responses were not measured in a condition where the signal was activated, we cannot observe this coordinated response.

In order to more fully decode transcriptomic networks, it is useful to measure RNA-seq responses in a wide variety of growth conditions. Many computational approaches have been developed to elucidate such networks from expression data[55] and as RNA-seq technology has matured, it has become a relatively accessible data type for researchers to collect, both in-house or through vendors.

2.2. *M. buryatense* RNA-seq datasets and relevant transformations

Over the past 40 years, the Lidstrom Lab (Microbiology and Chemical Engineering departments, University of Washington) has characterized various aspects of the physiology and genetics of methylotrophic bacteria that subsist on single-carbon compounds. During the past 10 years, over 100 RNA-sequencing experiments were executed in *M. buryatense* by the Lidstrom group, taking transcriptomic measurements across a variety of growth conditions, including ideal growth conditions (“max growth rate”), stress conditions (“low oxygen”, “low methane”), using an alternative carbon source (“methanol”), and in the presence of metals which are known to induce regulatory switches (“with lanthanum”, “high/medium/low copper”). While these experiments are quite varied and were designed with different goals in mind, the integration of all of these datasets presents the opportunity to investigate regulatory signals across a variety of conditions and infer subsets of genes that are influenced by similar regulatory signals. Much of the work in this dissertation stems from the goal to identify and characterize genetic signals in *M. buryatense* so they may be developed into engineering tools.

Some RNA-seq samples were experimental replicates, while others were gathered at different times under similar conditions. Every sample was assigned a broad experimental condition category, which was refined over the course of several quality control analyses. Additionally, several broad conditions were subdivided into more granular conditions after observing discrepancies in RNA-seq profile correlations (for example, “low oxygen” was split between “low oxygen, fast growth” and “low oxygen, slow growth”). These experimental condition labels are used primarily as a categorical encoding. While some experimental settings were more similar to each other than others, details about the exact growth settings are not available in every case and therefore not directly comparable on a relative continuum.

A description of each sample used throughout this dissertation and its corresponding experimental condition tag is available in Table 2.1. A summary of the number of samples in each condition is available in Table 2.2. In total, the *M. buryatense* data matrix contains 4,213 genes by 106 samples. After excluding samples with poor quality control or insufficient metadata about experimental conditions, we proceeded with 86 samples grouped into 17 experimental condition labels. Four samples for very low methane conditions (two replicates each for "CH_4 500 ppm" and "CH_4 1000 ppm") were collected partway through this dissertation - these samples were not available to use in the work described in Chapter 3 but were included for Chapters 4 and 5.

ID	sample name	description	experimental condition tag	rna-seq data published	5G strain	GEO link
0	5GB1_FM03_TR1_QC	Fermentor run 3, uMax though close to O2 limited, QC	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932954
1	5GB1_FM03_TR2_QC	Fermentor run 3, uMax though close to O2 limited, QC	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932955
2	5GB1_FM11_TR1_QC	Fermentor run 11, O2 limited, QC	lowO2_fast_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932956
3	5GB1_FM11_TR2_QC	Fermentor run 11, O2 limited, QC	lowO2_fast_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932957
4	5GB1_FM12_TR1	Fermentor run 12, methane limited	lowCH4	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720034
5	5GB1_FM12_TR1_QC	Fermentor run 12, methane limited, QC	lowCH4	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932958
6	5GB1_FM12_TR2	Fermentor run 12, methane limited	lowCH4	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720035
7	5GB1_FM12_TR2_QC	Fermentor run 12, methane limited, QC	lowCH4	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932959
8	5GB1_FM14_TR1	Fermentor run 14, methane limited	lowCH4	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720036
9	5GB1_FM14_TR1_QC	Fermentor run 14, methane limited, QC	lowCH4	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932960
10	5GB1_FM14_TR2	Fermentor run 14, methane limited	lowCH4	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720037
11	5GB1_FM14_TR2_QC	Fermentor run 14, methane limited, QC	lowCH4	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932961
12	5GB1_FM18_TR2	Fermentor run 18, methanol	MeOH	Fu 2019	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2995397
13	5GB1_FM18_TR2_QC	Fermentor run 18, methanol	MeOH	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932963
14	5GB1_FM18_TR3	Fermentor run 18, methanol	MeOH	Fu 2019	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2995398
15	5GB1_FM18_TR3_QC	Fermentor run 18, methanol	MeOH	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932964

ID	sample name	description	experimental condition tag	rna-seq data published	5G strain	GEO link
16	5GB1_FM19_TR1_QC	Fermentor run 19, O2 limited, QC	lowO2_fast_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932965
17	5GB1_FM19_TR1_UW	Fermentor run 19, O2 limited	lowO2_fast_growth	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720039
18	5GB1_FM19_TR3	Fermentor run 19, O2 limited	lowO2_fast_growth	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720040
19	5GB1_FM19_TR3_QC	Fermentor run 19, O2 limited, QC	lowO2_fast_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932966
20	5GB1_FM20_TR1_QC	Fermentor run 20, uMax, QC	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932967
21	5GB1_FM20_TR2_QC	Fermentor run 20, uMax	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932968
22	5GB1_FM20_TR3	Fermentor run 20, uMax	uMax	Fu 2019	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2995400
23	5GB1_FM20_TR3_QC	Fermentor run 20, uMax, QC	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932969
24	5GB1_FM20_TR3_UW	Fermentor run 20, uMax	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932970
25	5GB1_FM21_TR1	Fermentor run 21, uMax	uMax	Fu 2019	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2995401
26	5GB1_FM21_TR1_QC	Fermentor run 21, uMax, QC	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932971
27	5GB1_FM21_TR2	Fermentor run 21, uMax	uMax	Fu 2019	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2995402
28	5GB1_FM21_TR2_QC	Fermentor run 21, uMax, QC	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932972
29	5GB1_FM21_TR2_UW	Fermentor run 21, uMax	uMax	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932973
30	5GB1_FM22_TR1	Fermentor run 22, O2 limited	lowO2_fast_growth	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720041
31	5GB1_FM22_TR1_QC	Fermentor run 22, O2 limited, QC	lowO2_fast_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932974
32	5GB1_FM22_TR3_QC	Fermentor run 22, O2 limited, QC	lowO2_fast_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932975
33	5GB1_FM22_TR3_UW	Fermentor run 22, O2 limited	lowO2_fast_growth	Gilman 2017	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2720043
34	5GB1_FM23_TR3	Fermentor run 23, methanol	MeOH	Fu 2019	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2995399
35	5GB1_FM23_TR3_QC	Fermentor run 23, methanol	MeOH	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932976
36	5GB1_FM34_T0_TR1_QC	Fermentor run 34, Cu transition, T0 (before Cu) 5.15% FAME	NoCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932977
37	5GB1_FM34_T3_TR3_QC	Fermentor run 34, Cu transition, T3hr (after Cu) 5.49% FAME	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932978

ID	sample name	description	experimental condition tag	rna-seq data published	5G strain	GEO link
38	5GB1_FM34_T4_TR3_QC	Fermentor run 34, Cu transition, T4hr (after Cu) 5.12% FAME	highCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932979
39	5GB1_FM34_T5_TR2_QC	Fermentor run 34, Cu transition, T5hr (after Cu) 5.75% FAME	highCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932980
40	5GB1_FM34_T6_TR3_QC	Fermentor run 34, Cu transition, T6hr (after Cu) 5.71% FAME	highCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932981
41	5GB1_FM34_T7_TR3_QC	Fermentor run 34, Cu transition, T7hr (after Cu) 6.45% FAME	highCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932982
42	5GB1_FM34_T8_TR1_QC	Fermentor run 34, Cu transition, T8hr (after Cu) 6.20% FAME	highCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932983
43	5GB1_FM40_T0_TR1_QC	Fermentor run 40, Cu transition, T0 (before Cu)	NoCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932984
44	5GB1_FM40_T0m_TR2	Fermentor run 40, Cu transition, T0m before Cu)	NoCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932985
45	5GB1_FM40_T10m_TR3	Fermentor run 40, Cu transition, T10m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932986
46	5GB1_FM40_T10m_TR3_QC	Fermentor run 40, Cu transition, T10m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932987
47	5GB1_FM40_T150m_TR1_QC	Fermentor run 40, Cu transition, T150m (after Cu)	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932988
48	5GB1_FM40_T150m_TR1_remake	Fermentor run 40, Cu transition, T150m (after Cu)	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932989
49	5GB1_FM40_T180m_TR1	Fermentor run 40, Cu transition, T180m (after Cu)	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932990
50	5GB1_FM40_T180m_TR1_QC	Fermentor run 40, Cu transition, T180m (after Cu)	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932991
51	5GB1_FM40_T20m_TR2	Fermentor run 40, Cu transition, T20m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932992
52	5GB1_FM40_T40m_TR1	Fermentor run 40, Cu transition, T40m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932994
53	5GB1_FM40_T40m_TR1_QC	Fermentor run 40, Cu transition, T40m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932995

ID	sample name	description	experimental condition tag	rna-seq data published	5G strain	GEO link
54	5GB1_FM40_T60m_TR1	Fermentor run 40, Cu transition, T60m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932996
55	5GB1_FM40_T60m_TR1_QC	Fermentor run 40, Cu transition, T60m (after Cu)	lowCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932997
56	5GB1_FM40_T90m_TR2	Fermentor run 40, Cu transition, T90m (after Cu)	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932998
57	5GB1_FM40_T90m_TR2_QC	Fermentor run 40, Cu transition, T90m (after Cu)	medCu	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4932999
58	5GB1_FM69_t3_TR1	Fermentor run 69, high oxygen, slow growth rate	highO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933000
59	5GB1_FM69_t3_TR1_UW	Fermentor run 69, high oxygen, slow growth rate	highO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933001
60	5GB1_FM69_t4_TR1	Fermentor run 69, high oxygen, slow growth rate	highO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933002
61	5GB1_FM69_t4_TR1_UW	Fermentor run 69, high oxygen, slow growth rate	highO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933003
62	5GB1_FM80_t2_TR1	Fermentor run 80, extra nitrate, low oxygen, slow growth rate	NO3_lowO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933004
63	5GB1_FM80_t4_TR1	Fermentor run 80, extra nitrate, low oxygen, slow growth rate	NO3_lowO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933005
64	5GB1_FM81_t1_TR3	Fermentor run 81, extra nitrate, low oxygen, slow growth rate	NO3_lowO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933006
65	5GB1_FM81_t2_TR3	Fermentor run 81, extra nitrate, low oxygen, slow growth rate	NO3_lowO2_slow_growth	Wilson 2021	5GB1	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933007
66	5GB1_vial_wLa_TR3	Vial sample, with lanthanum	WithLanthanum	Wilson 2021	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933008
67	5GB1_vial_woLa_TR2	Vial sample, without lanthanum	NoLanthanum	Wilson 2021	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4933009
68	5GB1C-5G-La-BR1	Vial sample in mid- to late-exponential phase	WithLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584843
69	5GB1C-5G-La-BR2	Vial sample in mid- to late-exponential phase	WithLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584844
70	5GB1C-5G-N-BR1	Vial sample in mid- to late-exponential phase	NoLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584845
71	5GB1C-5G-N-BR2	Vial sample in mid- to late-exponential phase	NoLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584846

ID	sample name	description	experimental condition tag	rna-seq data published	5G strain	GEO link
72	5GB1C-JG15-La-BR1	Vial sample in mid- to late-exponential phase	WithLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584847
73	5GB1C-JG15-La-BR2	Vial sample in mid- to late-exponential phase	WithLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584848
74	5GB1C-JG15-N-BR1	Vial sample in mid- to late-exponential phase	NoLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584849
75	5GB1C-JG15-N-BR2	Vial sample in mid- to late-exponential phase	NoLanthanum	Groom 2019	5GB1C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3584850
76	5GB1_LTrecycle_TR1	LanzaTech Cell recycle	LanzaTech	NA	5GB1	NA
77	5GB1_LTrecycle_TR1_QC	LanzaTech Cell recycle	LanzaTech	NA	5GB1	NA
78	5GB1_FM_85_TR1	Fermentor run 85 with aa3 knockout	aa3_KO	NA	5GB1	NA
79	5GB1_FM_85_TR2	Fermentor run 85 with aa3 knockout	aa3_KO	NA	5GB1	NA
80	5GB1_pA9_red	Vial sample producing crotonic acid	crotonic_acid	NA	5GB1	NA
81	5GB1_pA9_yellow	Vial sample producing crotonic acid	crotonic_acid	NA	5GB1	NA
82	5GB1C_CH4_500ppm-Rep1	Fermentor run with CH4 concentration at 500 ppm	CH4_500ppm	He (under review)	5GB1C	NA
83	5GB1C_CH4_500ppm-Rep2	Fermentor run with CH4 concentration at 500 ppm	CH4_500ppm	He (under review)	5GB1C	NA
84	5GB1C_CH4_1000ppm-Rep2	Fermentor run with CH4 concentration at 1000 ppm	CH4_1000ppm	He (under review)	5GB1C	NA
85	5GB1C_CH4_1000ppm-Rep1	Fermentor run with CH4 concentration at 1000 ppm	CH4_1000ppm	He (under review)	5GB1C	NA

Table 2.1. Complete list of RNA-seq samples used in this dissertation.

experimental_condition_tag	sample_count
uMax	12
lowO2_fast_growth	10
lowCH4	8
medCu	7
lowCu	7
MeOH	6
WithLanthanum	5
highCu	5
NoLanthanum	5
highO2_slow_growth	4
NO3_lowO2_slow_growth	4
NoCu	3
LanzaTech	2
crotonic_acid	2
aa3_KO	2
CH4_500ppm	2
CH4_1000ppm	2
Total	**86**

Table 2.2. Summary of RNA-seq sample counts per experimental condition.

To process the raw RNA-seq data into the data matrix mentioned above, standard bioinformatics tools were used, assisted by the barrelseq workflow (available on Github: https://github.com/BeckResearchLab/barrelseq). Briefly, reads from each sample fastq file were aligned to the *M. buryatense* genome (NCBI accession NZ_CP035467.1) using BWA with the BWA-MEM algorithm (BWA version 0.7.17-r1198-dirty, default parameters).[56] SAMTools version 1.9 was used to transform the initial read alignments into sorted BAM files.[57] The htseq-count tool from the HTSeq framework version 2.0.2 was used to attribute the reads to annotated features using the “intersection-nonempty” mode, providing estimates of raw read counts.[58] Raw read counts were subsequently converted into transcripts per million (TPM) for each genome feature in order to normalize the counts by the feature length.[59]

Transformations of TPM values are the primary data type used throughout the analyses in this dissertation. In most cases, the TPMs were $\log_2$ transformed and in some analyses, a log ratio was calculated relative to a baseline condition.

2.3. Initial dataset characterization and exploration

To characterize some initial features of these data, we briefly looked into the overall TPM range of all the genes averaged within each experimental condition and noted that all experimental conditions show similar ranges of TPM values. Distributions of expression values within each experimental condition indicated that most genes express at a relatively low level (between 20 to 100 TPM) while a handful are orders of magnitude higher (Figure 2.1.A).

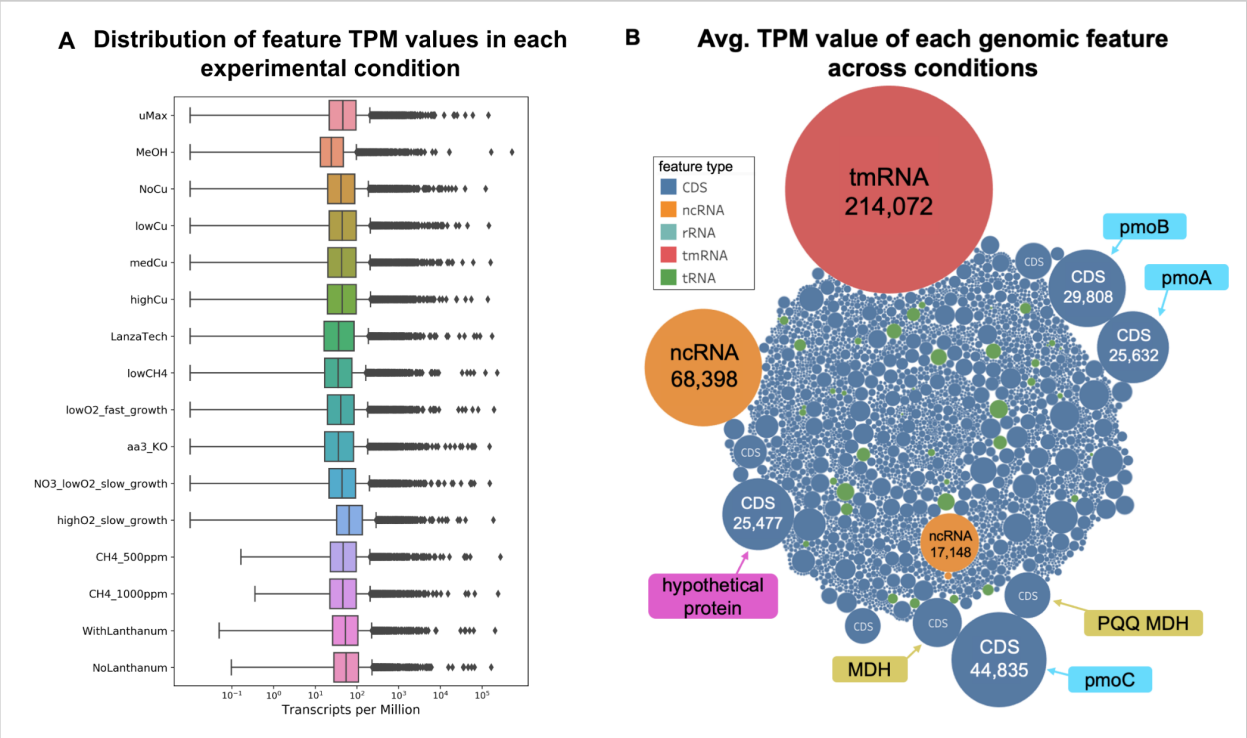


Figure 2.1. Distribution of TPM values. **A)** Gene transcript per million (TPM) values were averaged within each experimental condition. The distribution of TPM values is plotted on a log scale. **B)** Gene TPM values were averaged across all conditions. Several non-coding RNA features express at extremely high magnitudes, above even the pMMO gene cluster, which had not been previously known in *M. buryatense*.

Unsurprisingly, genes in the pMMO gene cluster (*pmoC*, *pmoA*, *pmoB*), which is involved in the first step of the *M. buryatense* methane assimilation pathway and known to express very strongly, were among these outliers. Upon further investigation, we noticed that the most highly expressed features were an ncRNA (EQU24_RS19765) and a tmRNA (EQU24_RS12525). Briefly, ncRNAs do not code for a protein but instead typically have a regulatory function[60] while tmRNAs are specialized to have both “transfer” and “messenger” RNA activity which allow them to help recycle ribosomes that have stalled on defective transcripts.[61] In *M. buryatense*, the ncRNA and the tmRNA were consistently expressed higher than all coding sequence (CDS) features -

even the suite of *pmo* genes (Figure 2.1.B). As previous analyses of RNA-seq data in *M. buryatense* primarily focused on protein-coding sequences, these trends were not previously known and warrant further investigation.

We next visualized relatedness among RNA-seq samples in *M. buryatense* gene expression space. Specifically, we performed principal component analysis (PCA) to reduce gene-dimensional space (~4000 genes) to two. Broadly, samples assigned the same experimental condition label visibly co-locate (Figure 2.2.A.) and 90% of the variance is explained by the first 12 components (Figure 2.2.B).

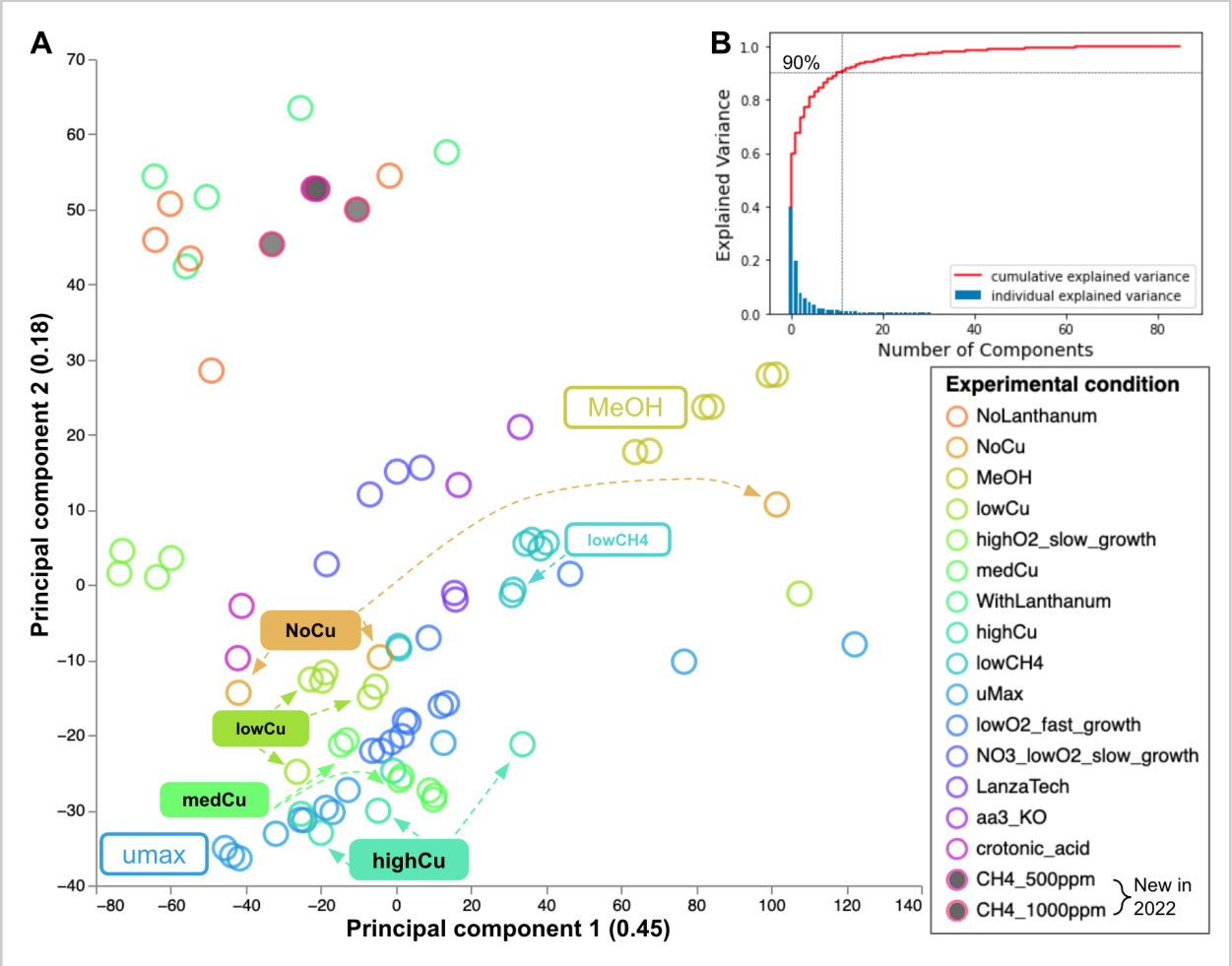


Figure 2.2. Principal Component Analysis of RNA-seq samples. **A)** Each circle is an experimental sample, colored by its condition tag. Annotations: samples involved in a Copper transition experiment are roughly captured in the relative orientation of “No copper”, “Low copper”, “Medium copper”, and “High copper” samples along a consistent axis. “High copper” samples are more similar to those in ideal growth conditions (uMax) experiments while “No copper” samples are more similar to methane limited (lowCH4) experiments in components 1 and 2, which account for 0.45 and 0.18 of the explained variance, respectively. **B)** A cumulative sum of the variance explained by the principal components. 90% of the variance is explained by the first 12 components.

Notably, the four copper-related conditions (a copper transition experiment yielding “no copper”, “low copper”, “medium copper” and “high copper” experimental tags) appear to align roughly in order of copper concentration. In principal components 1 and 2, “high copper” (highCu) samples are co-located near “max growth rate” (uMax) samples while lower copper (noCu, lowCu) skew towards “low methane” (lowCH4) and “methanol substrate” (MeOH) samples, suggesting that perhaps this dimension captures some overall growth rate or stress trend. This is consistent with the fact that copper is necessary for the particulate methane monooxygenase genes (*pmoA*, *pmoB, pmoC*) to be expressed, as opposed to the alternative soluble form of the enzyme which is instead dependent on iron. As methane monooxygenase genes encode the enzyme responsible for the first stage of methane assimilation into central metabolism, the similarity of samples in ideal growth conditions to those in high copper (preferred methane assimilation pathway) conditions is expected.

Another striking trend is that the “With Lanthanum” and “Without Lanthanum” conditions are noticeably more separated from the main axis of variation. Given that these conditions were intended to show contrast between Lanthanum-regulated genes, their increased similarity to each in reduced dimensions was unexpected. We initially suspected that it is likely due to their unique experimental setup in vials rather than bioreactors, which would exert somewhat different environmental growth pressures on the bacteria. After appending four additional samples belonging to the “CH_4 500 ppm” and “CH_4 1000 ppm” experimental tags, we saw these samples also co-locate with the lanthanum vial experiments despite being grown in bioreactors (filled circles in 2.2.A). One potential bias causing this similarity could be the recency of data collection of these samples: the lanthanum-vial and CH_4 500/1000 ppm samples were all collected within the last 5 years during which measurement technology has progressed. In Figure 2.1.A, we see that the lower range of TPMs for the lanthanum-vial and CH_4 500/1000 ppm conditions is slightly higher than the rest. Perhaps the first two principal components are capturing a discrepancy in which several genes that received 0 read counts in past sequencing runs were better detected in more recent sequencing runs due to reduced dropout error.

A visualization enabling interactive exploration of the first four principal components is available here: erinhwilson.github.io/interactive-thesis/viz_pages/chapter2_PC1234.html

2.4. Strengths and limitations as a representative dataset

This compendium of RNA-seq samples serves as a representative dataset that many labs working with non-model organisms may be able to compile. RNA-sequencing technology is no longer a highly-specialized measurement to collect in the lab and is

readily available through popular vendors such as Illumina and Genewiz. Additionally, many computational tools for read alignment and counting are available open-source (BWA[56], Samtools[57], htseq-count[58], the STAR aligner[62]), enabling any researcher with an annotated genome to analyze their organism's transcriptome data. Several labs have recently taken advantage of their accumulated RNA-seq datasets collected over the course of other publications and further analyzed them together.[63–68] Developing computational tools that can readily combine old datasets and discover additional insights encourages data reuse and FAIR (Findable, Accessible, Interoperable, Reusable) accessibility practices.[69]

Chapters 3 and 4 of this dissertation discuss computational approaches that can leverage varied RNA-seq datasets for investigating microbe regulatory patterns. We apply them in the context of the *M. buryatense* RNA-seq dataset, however the frameworks are built to be generalizable to other microbes.

A limitation of this type of RNA-seq dataset is that the number of gene examples is limited by the genome of the organism of interest. Deep learning approaches (discussed in Chapter 5) typically rely on large training datasets in order to accurately learn model weights relevant to complex prediction tasks and often yield better performance as the dataset size increases.[70,71] And indeed, there exists excellent work developing models to predict diverse gene expression behaviors! Several approaches leverage Massively Parallel Reporter Assay (MPRA) data: datasets composed of hundreds-of-thousands to millions of random sequence examples, their influence on gene expression measured in high throughput experiments.[72–77] Other works in eukaryotes take advantage of measurement types such as ChIP-seq or ATAC-seq, where the number of example sequences is not directly tied to the number of genes but can detect an arbitrarily large number of peak instances measured throughout the regulatory regions of a genome.[78–81]

At the outset of the project discussed in Chapter 5, we were unsure if the 4,000-feature genome of *M. buryatense* would be sufficient for a model to learn the complexity of its regulatory grammar, as we did not have MPRA data or peak data for this organism. Intrigued by the possibility, we investigated the potential of using RNA-seq data for decoding regulatory motif patterns, but our results indicate that it is likely insufficient on its own. We further analyzed synthetic sequence prediction tasks in variable data-limited regimes and discuss the level of dataset information richness necessary to capture a given degree of motif signal complexity. This characterization of deep learning model performance in relatively small microbial genomes will be informative for other researchers planning machine learning-based investigations of similarly-sized transcriptomic datasets and we suggest ways to incorporate additional data that may lead to more fruitful results.

Overall, initial investigations of TPM expression distributions and PCA visualizations of the *M. buryatense* RNA-seq compendium emphasized that this varied dataset captures known physiological trends and thus is a promising source to mine for further biological insights while calling out technical caveats for consideration in downstream analysis. Future analyses may benefit from additional normalization strategies to account for potential differences in updated technologies and reduce batch effects.

Chapter 3. A computational framework for identifying promoter sequences in non-model organisms using RNA-seq datasets

3.1. Background and related work

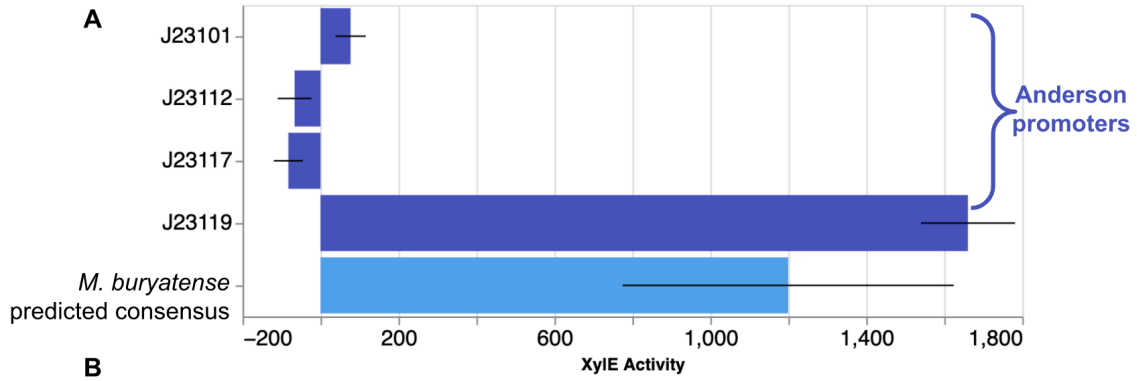
For any choice of microbial host, a successful biomolecule production platform relies on the ability to optimize an organism's metabolism to produce a target molecule efficiently and at high yields, which in turn relies on our ability to precisely control gene expression.[5,6] In prokaryotes, the -35 and -10 hexamers in the promoter are particularly key for transcription initiation.[20,82] Testing previously developed promoter tools, such as those developed in well-studied model systems like *E. coli*, is a useful place to start when adopting a new host. While the sigma-70 transcription factor is a well-conserved initiator of housekeeping genes across many prokaryotes, different microorganisms might have altered motif preferences.[27,83] When readily available tools are insufficient, it prompts the need to develop a new framework for building out genetic toolkits explicitly tailored to new organisms.

A number of previous studies have proposed promoter prediction software based on a variety of computational techniques, such as expectation maximization, kernel alignment, hidden markov models, DNA stability, and neural networks.[84–88] However these efforts are primarily focused on *E. coli* promoters and validation relies on pre-curated data sets[82] or databases of known promoters.[89–91] For non-model organisms, which lack such databases, promoters must first be identified, annotated, and experimentally characterized before such techniques can be applied. In order to precisely identify promoters, there exist specialized RNA-seq protocols, such as differential RNA-seq[92] and 5'-RACE,[93] that aid in the identification of transcription start sites (TSSs). However, not all labs routinely collect these specialized data. Thus, a method that relies solely on common data types - such as whole genome and RNA-sequencing - would be beneficial in expanding the range of existing information that could be utilized for developing promoter tools.

Over the past decade, the Lidstrom lab at the University of Washington has collected RNA-seq samples measuring gene expression across a variety of conditions in the methanotroph *M. buryatense* (more details available in Chapter 2). *M. buryatense* is an extremely relevant organism for cycling methane, a potent greenhouse gas whose emissions contribute to more than 20% of anthropogenic climate change.[40] Previous work in *M. buryatense* developed genome editing tools, a full-scale metabolic model, and characterized a lanthanide metal switch.[27,43,45,94–96] Notably, Puri *et al.* confirmed that

E. coli promoter sequences such as P_{tac} and P_{lac} can drive the expression of a reporter gene; however, the ranking in expression strength of these promoters relative to the *M. buryatense* native P_{mxaF} was not preserved between organisms.[27]

Following up on this observation, we performed a preliminary analysis exploring the potential for the suite of Anderson promoters[97] developed for *E. coli* to work in a methanotroph. While these sequences show strong to moderate expression in *E. coli*, we found them to show differing expression strengths when used in *M. buryatense*: only construct J23119 exhibited strong expression in *M. buryatense* while J23101, J23112, and J23117 showed little to no expression in this methanotrophic host (Figure 3.1.). This suggests that the regulatory grammar underlying *M. buryatense* is different enough that solely using *E. coli*-based expression tools would be insufficient to effectively manipulate gene expression in this organism and additional work to build up an expression toolkit is needed. Given the compendium of RNA-seq data available and this organism's promising potential to serve as a metabolic engineering platform, we sought to build an RNA-seq-based computational framework to develop promoter tools.



B

Promoter name	*M. buryatense* expression	*E. coli* expression	Sequence	Plasmid
J23101	very low	high	TTTACAgctagctcagtcctaggTATTAT	pSMF7
J23112	none	moderate	CTGATAgctagctcagtcctaggGATTAT	pSMF3
J23117	none	moderate	TTGACAgctagctcagtcctaggGATTGT	pSMF2
J23119	high	high	TTGACAgctagctcagtcctaggTATAAT	pSMF8
Predicted consensus	high	N/A	TTGACAaggtctattaggcagtgTATAAT	pMCS37.2

Figure 3.1. Anderson promoter expression in *M. buryatense*. **A)** XylE reporter assay results for four Anderson promoters compared to the predicted consensus promoter for *M. buryatense*. The x-axis represents XylE reporter activity in milli-Units / min normalized to OD_{600}. Error bars represent the standard deviation of 3 technical replicates. **B)** Sequence and expression information of the Anderson promoters when used in *E. coli* versus *M. buryatense*. Notably, J23101, J23112, J23117 all exhibit moderate to high expression in *E. coli* but are very low or non-detectable in *M. buryatense*. J23119 is

the only Anderson promoter to show high activity in both *E. coli* and *M. buryatense* and shares the exact -35 and -10 hexamer as the predicted consensus for *M. buryatense*.

Several recent efforts have used RNA-seq data to survey microbial genomes for constitutively, strongly expressed genes.[65–68] For example Luo *et al* [67] discovered a panel of 25 constitutive promoters in *Streptomyces albus* while Ouyang *et al* [68] similarly characterized 37 for a Burkholderiales strain. However, many of the promoter sequences reported in these methods are hundreds of nucleotides long or encompass the entire upstream window between the translation start site and the next upstream gene. Though using a large upstream window ensures that the extracted sequence contains the promoter elements involved in transcription initiation, these regions may also contain other regulatory signals that, if incorporated in a heterologous expression cassette, may result in unanticipated expression effects. Identifying minimal promoters that contain only the core transcription initiation signal would 1) improve predictability by isolating the signal from its surrounding context, and 2) reduce the genetic manipulation burden of working with expression constructs containing long stretches of genome homology that may induce unintended recombination events.[98,99]

3.2. A computational framework for promoter identification

To enable the development of genetic tools that are both explicitly tailored to a new organism of interest and composed of relatively short DNA sequences that still confer strong expression, we present the following computational framework (Figure 3.2.). It proceeds in three main stages: 1) identification of a group of highly expressed genes that maintain high transcript counts across a broad range of experimental conditions, 2) extraction of the corresponding upstream candidate promoter regions of these highly expressed genes while avoiding regions upstream of genes that may reside in operons, and 3) application of the motif-finding algorithm in BioProspector[100] to these upstream regions to predict the location and sequence of the -35 and -10 hexamers that drive the strong expression of these loci.

To identify sets of constitutively, highly expressed genes in stage 1, the TPM expression values for each gene were calculated as an average within each of the experimental conditions represented in the *M. buryatense* RNA-seq compendium. After ranking genes in each condition from high to low, candidate sets of highly expressed genes (referred to as “top genes”) were constructed at varying thresholds *n* such that all genes in a candidate set remained in the top *n*% of genes across all conditions.

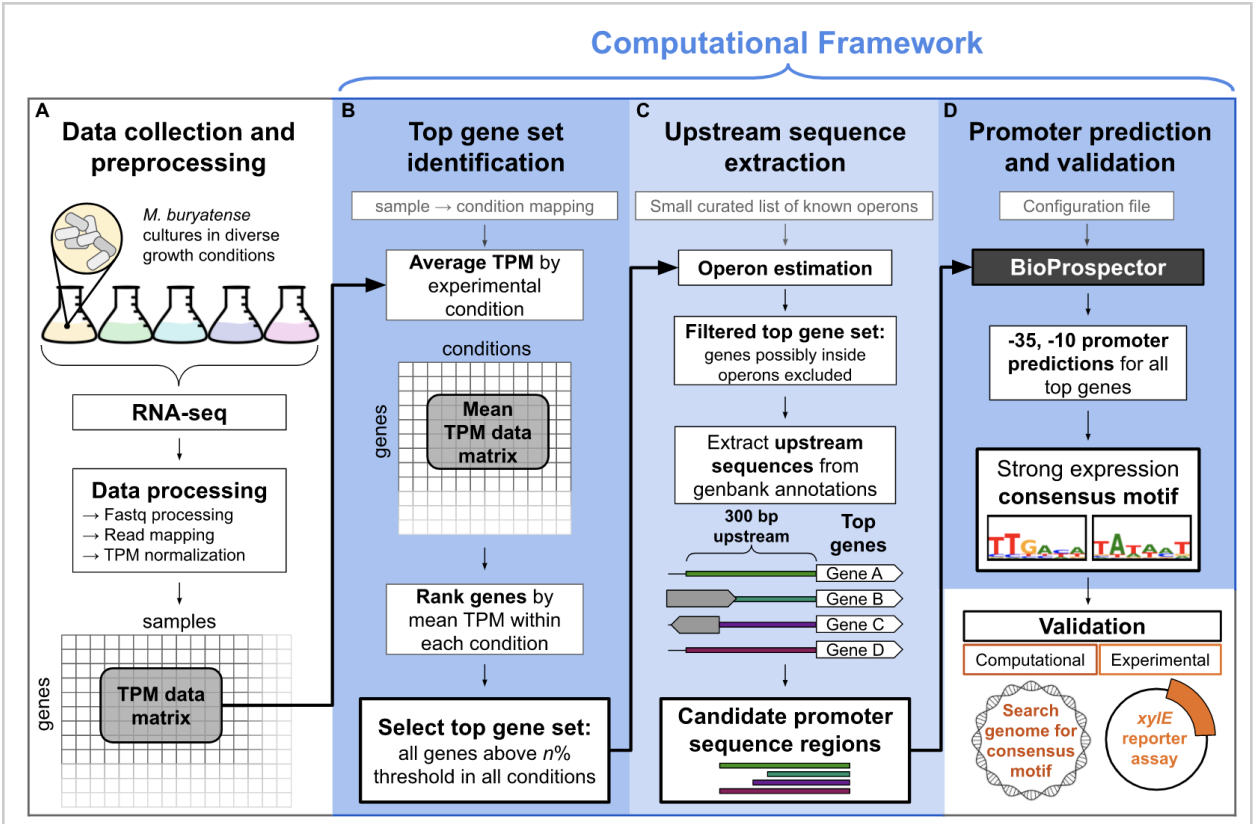


Figure 3.2. Overview of promoter prediction framework. **A)** Independently, RNA-seq data are collected and processed to reflect TPM counts. The framework proceeds in 3 main stages. **B)** First, samples are aggregated by their assigned experimental conditions and averaged by gene. Top gene sets are identified by ranking all genes from high to low expression and keeping all genes that appear in the top *n*% of every condition for a given value of *n*. **C)** Before extracting upstream sequences, some of these genes are filtered out if they are predicted to reside *inside* an operon (any operon-affiliated gene except the first gene). Using this filtered set of top genes, we access their genbank annotations and extract a sequence window 300 bases upstream from the feature start coordinate, though this window is truncated to include only intergenic DNA if a neighboring feature appears within 300 nucleotides. **D)** Finally, these upstream regions are fed to the BioProspector motif discovery tool to predict the best -35 and -10 promoters for each top gene. These predictions are additionally compiled into a consensus motif for strong expression and are validated both computationally and experimentally.

After examining candidate top genes sets at every threshold between 1 and 10, we observed a steep tradeoff between the number of top genes belonging to a set and the lower bound of TPM expression among members of the top set (Figure 3.3.A). We chose to proceed with the set corresponding to the top 3%. This balanced the number of genes that qualify as being in the top expressed set (37 genes) while ensuring that the average expression of members in the set are indeed strong - all members of the top 3% set have consistently higher TPM expression values than the first standard deviation of the remaining 97% of genes in every experimental condition. Most members of the top 3% set maintain stronger expression than even the second standard deviation of remaining genes (Figure 3.3.B).

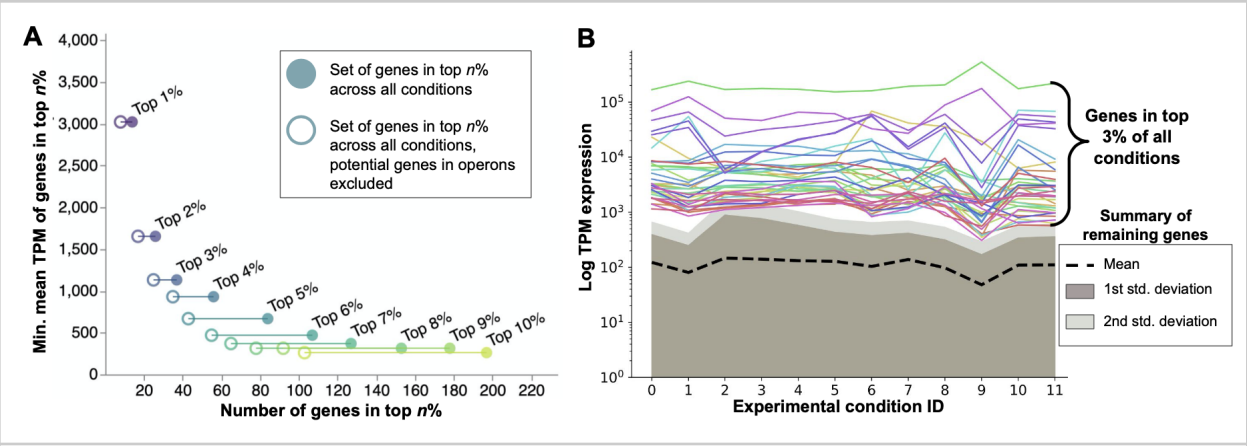


Figure 3.3. Identification of top gene sets. **A)** Tradeoff between gene count and TPM expression for different top gene sets depending on the *n*% threshold. Each point denotes a top gene set and shows the total count of all genes in the set on the x-axis versus the minimum gene's mean TPM value on the y-axis. For small values of *n*, the mean TPM expression of the weakest gene in the top set increases, while the total count of genes in the top set decreases, and vice versa. Filled circles represent all genes that qualify as the top *n*% across all conditions. Open circles represent filtered versions of the same set, where any gene predicted to be inside an operon (any except the first gene) has been excluded. **B)** Expression strength of the top 3% gene set compared to the remaining genes in the genome. Each colored line represents the TPM expression (log-scale) trend of a top gene across each of the conditions in the RNA-seq compendium. The black dashed line represents the average log TPM expression of all remaining genes in the genome. The dark and light gray bands represent the first and second standard deviations, respectively.

The next stage of the framework extracts the sequences immediately upstream of each top gene. These sequences are likely to contain promoter regions with constitutive regulatory signals and the extraction was easily accomplished using genbank annotation files and biopython software tools. However, since promoters are primarily expected upstream of genes transcribed singly or upstream of the first gene in an operon,[101] it is important to exclude top genes that are likely to be downstream genes residing inside operons; their immediate upstream sequences are likely to be the coding sequences of neighboring genes rather than promoter regions.

Like many non-model bacteria, the operons of *M. buryatense* have not yet been formally annotated. To exclude genes likely to be inside operons from the pattern identification phase of the framework, we implemented a simple operon estimation strategy. Specifically, we used BioPython to locate the feature annotation of each gene in the *M. buryatense* genbank file (NCBI accession: CP035467.1) and scanned upstream to the next feature located on the same strand. If the neighboring feature was within 120 bases of the candidate top gene, the top gene was excluded from further promoter analysis. We chose 120 based on analysis of intra-operon gene distances of a small list of known operons in *M. buryatense*: at this threshold, the bulk of known operon clusters were captured, including the highly expressed pMMO gene cluster (Figure 3.4.).

However, this operon distance setting is a flexible parameter in the computational framework and can be configured by the user. Adjusted counts of top gene candidate sets with possible operon genes excluded are visualized as open circles in Figure 3.3.A.

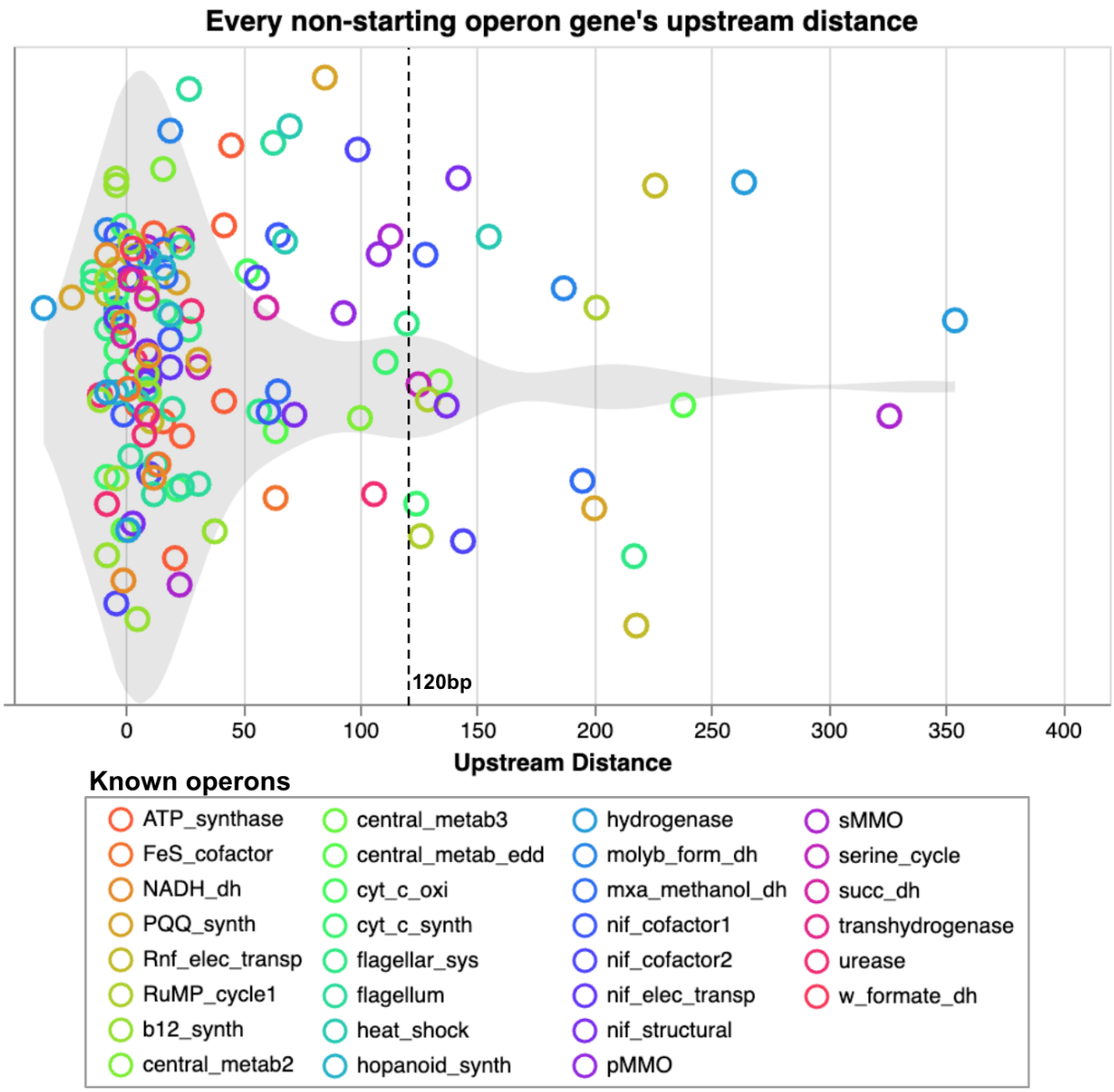


Figure 3.4. *M. buryatense* intra-operon upstream distances. Each point is a gene within a known operon, colored by its operon membership. The x-axis is the distance in base pairs to the gene's immediate upstream neighbor within the same operon. The y-axis is jittered to spread out the points. A violin plot underlay shows the distribution of genes' intra-operon distances. A threshold of 120bp is used as the estimate for genes with unknown operon memberships.

Sequences extracted in stage 2 were provided as input to the motif detection tool BioProspector,[100] which was configured to search for a -35, -10 motif structure: a pair of

hexamers separated by a 15-18bp spacer (Figure 3.5.A). Many computational tools have been proposed for motif discovery tasks;[102] we chose BioProspector for its usability and flexible settings that enabled searching for 2-block motifs with variable spacing (the structure of most known strong promoters in bacteria). The BioProspector algorithm searches all input sequences for a common motif pattern that fits the requested structure and reports five motif predictions, which are not necessarily unique. Additionally, it reports the coordinate of the subsequence within each input sequence where each motif match was found.

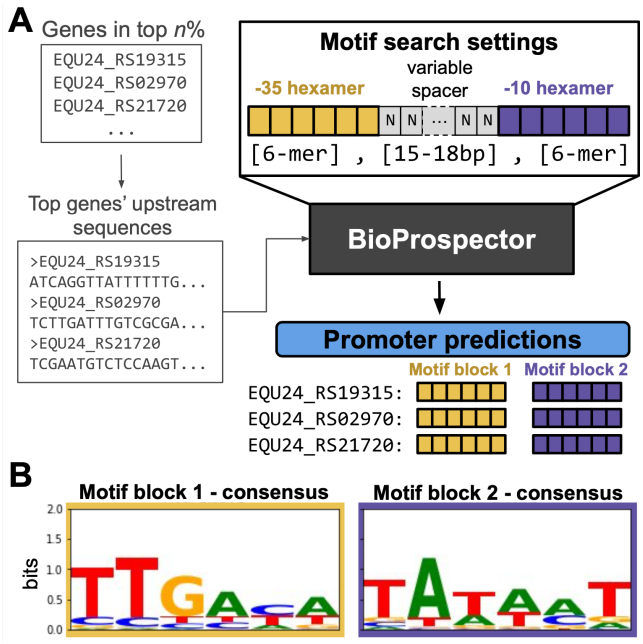


Figure 3.5. Building a consensus motif from promoter predictions. **A)** Genes in the top *n*% gene set have their upstream sequences extracted from a genbank file and stored in a fasta file. These are fed to BioProspector, which searches among the input sequences for a common motif pattern. By default, BioProspector is configured to search for a pair of hexamers separated by 15-18 nucleotides. After compiling 200 *in silico* replicates of the BioProspector search results, the top predictions for the -35, -10 promoter region are reported. **B)** -35 and -10 consensus motifs are constructed from the promoter predictions by aligning the first six and final six bases, respectively, of each prediction. The resulting PWMs are visualized as a sequence logo using the python package Logomaker.

Due to the stochastic nature of its search process, our framework executes multiple *in silico* replicates of the BioProspector search process (200 by default), yielding many motif predictions. For each promoter input, we count the total votes for each promoter candidate (i.e., the number of times BioProspector identifies the exact same subsequence within the input promoter region as matching a predicted motif from any of the *in silico* replicates). The most popular subsequence in terms of BioProspector votes is selected as the best promoter prediction for each gene. Predictions are also reported along with the *margin of victory*, that is, the difference in votes received by the predicted sequence and the next most popular sequence. The margin of victory value helps to distinguish cases where BioProspector very consistently predicted the same sequence as the promoter signal, resulting in a high margin of victory, versus cases where multiple potential promoter candidates were identified and the BioProspector motif search found multiple candidates with similar frequency, which would result in a low margin of victory. Users have access to the top *k* promoter predictions. The value for *k* is 3 by default but

may be specified by the user. Margins of victory are reported for each input sequence, letting users tap additional biological expertise or insight to override any predictions.

For each input gene, the sub-sequence that received the most votes was called as the best -35, -10 promoter candidate. These top -35 and -10 predictions were compiled into consensus hexamer motifs by creating position weight matrices (PWMs) from the first six and final six bases, respectively, of the top-predicted promoter candidate sequences for all inputs and visualized with Logomaker.[103] We found the consensus motif to match the canonical housekeeping consensus in *E. coli*: TTGACA, TATAAT (3.5.B). This is partially expected given that previous work in *M. buryatense* was able to use the *E. coli* P_{tac} and P_{lac} promoters to drive reporter gene expression, albeit less strongly than the native P_{mxaF}.[27]

3.3. Computational validation: consensus motif frequency by genomic region

To determine whether the consensus motif we discovered was correlated with *M. buryatense* promoter regions, we searched through the entire *M. buryatense* genome for sequence matches to the -35, -10 consensus motif determined from the top 3% of genes (Figure 3.6.). Matches to the consensus were determined by computing four variably-spaced position specific scoring matrices (PSSM) using the BioPython[104] Motif module (Figure 3.6.A). We then calculated the log odds score that each genome subsequence matched one of the consensus PSSMs. Each match greater than 0.0 was assigned to one of four genome categories based on the match position: 1) inside an annotated feature, 2) intergenic, farther than 300 bases from a feature start coordinate, 3) intergenic, within 300 bases of a feature start coordinate, 4) intergenic, within 100 bases of a feature start coordinate (Figure 3.6.B).

The consensus motif was found in all four genome categories; however, after normalizing the number of motif matches found by the number of possible genome positions in each category, we found a higher frequency of consensus PSSM matches in regions immediately upstream of genes. The highest concentration of matches occurred in the 100 bases immediately upstream of an annotated feature (Figure 3.6.C). When considering only consensus motif matches of very high quality (log odds > 12.0), this enrichment was even more pronounced (Figure 3.6.D). These results support the assertion that the motif signal we identified is correlated with promoter regions and was not erroneously detected.

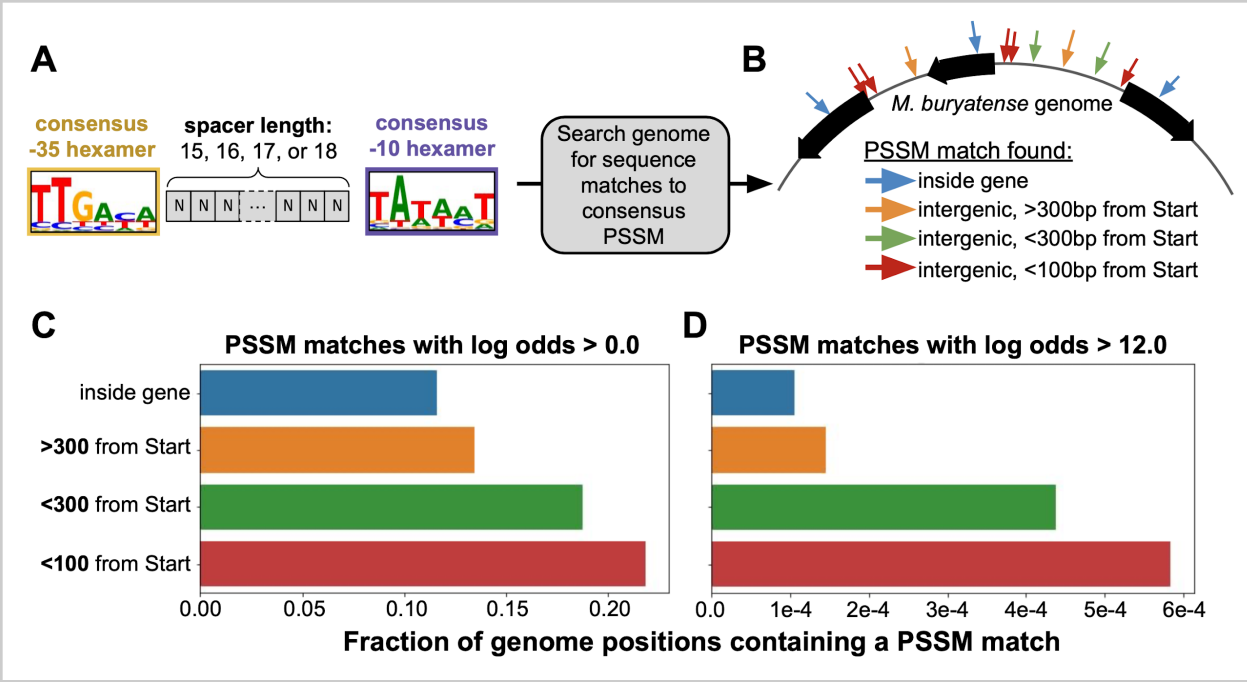


Figure 3.6. Computational validation of the consensus motif. **A)** Four position specific scoring matrices (PSSMs) are created from the -35 and -10 consensus motifs with intervening spacers of 15, 16, 17, or 18 nucleotide positions. All four PSSMs are used to search the entire genome sequence on both strands for sequence matches to the PSSM. Any sequence with a log odds score above 0.0 (greater similarity to the PSSM than random) is recorded as a match. **B)** Each match is assigned to one of four genome region categories based on the position of the match. The four regions are 1) inside a gene feature annotation, 2) intergenic, beyond 300 nucleotides from an annotated feature start coordinate, 3) intergenic, within 300 nucleotides from an annotated feature start coordinate, or 4) intergenic, within 100 nucleotides of an feature start coordinate. Category 4 is a subset of Category 3. **C)** The count of consensus PSSM matches in each genome region is normalized by the total number of genomic positions in each category and plotted to depict the fraction of positions in each category where a match was found. Promoter regions (categories 3 and 4) are enriched for sequences similar to the consensus motif. **D)** The set of matches was filtered to include only extremely high quality matches (sequences that matched consensus PSSM with a log odds > 12.0).

We repeated this analysis for different values of *n* to examine how the decision to use a gene set from varying top *n*% thresholds might influence consensus motif prediction results (Figure 3.7). Unsurprisingly, the consensus motif determined from only the top 1% of constitutively expressed genes was non-specific to promoter regions - it appeared equally frequently in all four genome categories - and likely a false result due to small sample size. At the other extreme, the consensus motifs found from the top 10% and 20% gene sets showed increased frequency in promoter regions, albeit with far fewer high quality matches of the consensus. This result is probably because the 10% and 20% consensus motifs had an overall lower information content, as the signal for “strong expression” was likely muddled by genes included in the top gene set but were not actually expressed very strongly. Consensus motifs from the top 2% and 5% sets were similar to the 3% set discussed earlier and showed large enrichment of high

quality matches (log odds > 12.0). However, the top 4% set differed significantly: it had much lower information content and yielded no high quality sequence matches despite only minor differences in gene set membership. After further investigation, we are not certain of the underlying reason for this deviation but suspect that the BioProspector algorithm may be sensitive to the introduction of certain subsets of genes that subtly pull the motif detection towards a weaker, competing signal. We thus recommend that future analyses with this framework test multiple thresholds of *n* and proceed with a robust consensus motif that shows enrichment of high quality sequence matches in promoter regions.

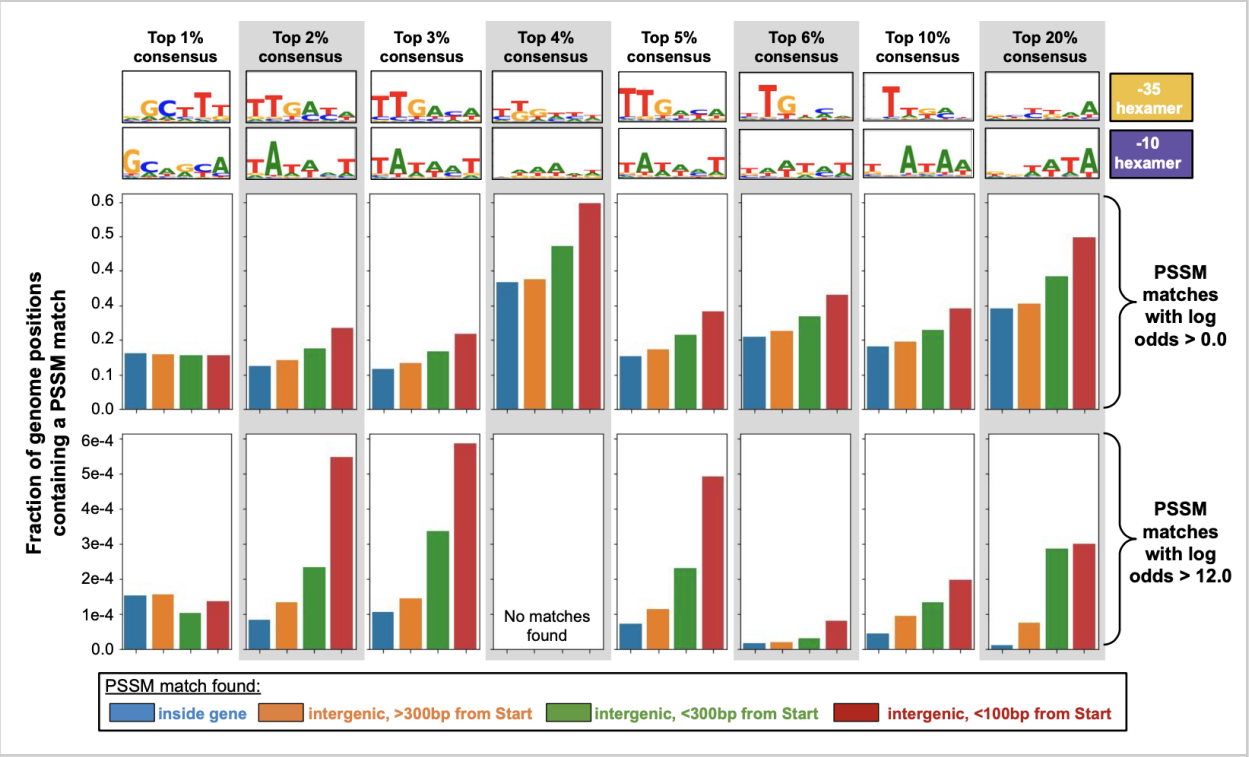


Figure 3.7. Consensus motifs derived from varying top *n*% gene sets. Consensus motifs were determined from upstream sequences of genes and used to search the entire genome for matches. Matches were assigned to genome categories based on their positions and normalized by the number of positions in each genome region, as in Figure 3.6. The top row of bars shows enrichment for all matches (log odds > 0.0) while the bottom row of bars shows enrichment for only very high quality PSSM matches (log odds > 12.0).

3.4. Experimental validation: assessing transcriptional activity of promoter predictions via a xylE reporter assay

To further validate the biological relevance of promoter predictions from this computational workflow, we next evaluated the predictions using a *xylE* reporter assay

to measure transcriptional activity (Figure 3.8.A). Six promoter predictions, each only 27-30 nucleotides in length, demonstrated XylE activity greater than the promoter for *rpoD*, a gene previously used as a reference for stable, moderate expression in *M. buryatense*[96] (Figure 3.8.B). We additionally created a synthetic construct containing the exact -35 (TTGACA) and -10 (TATAAT) consensus hexamers with a random 17bp spacer. This synthetic promoter does not appear anywhere in the *M. buryatense* genome but it demonstrated strong XylE activity, with only one of the six native promoter predictions exhibiting higher activity (Figure 3.8.B). The minimal consensus promoter sequence generated here is a new tool for driving high constitutive expression in this bacterium, and the other promoters identified can be used for increased or decreased expression relative to the consensus.

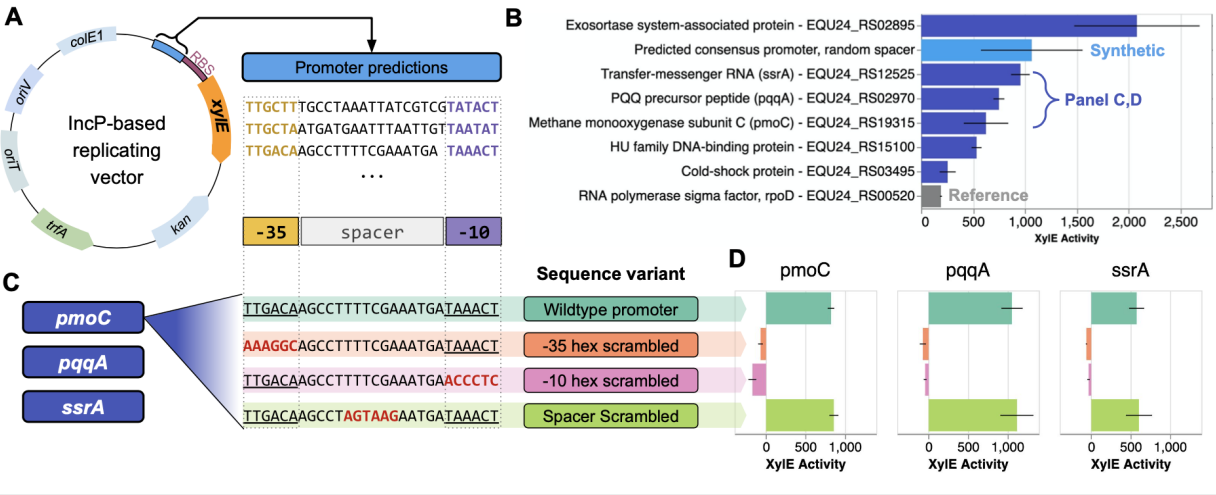


Figure 3.8. Experimental validation of promoter predictions. **A)** Vector map of replicating plasmids used in reporter gene assay. **B)** Expression results for 8 promoter constructs: 6 are members of the top 3% promoter prediction set (dark blue); one construct was synthetically assembled to contain the exact consensus for the predicted -35 (TTGACA) and -10 (TATAAT) motifs with a random spacer of length 17 (light blue); and the final construct is the promoter for *M. buryatense* RNA polymerase sigma 70 *rpoD* (gray), which was used a representative reference for stable, moderate expression in *M. buryatense*. The x-axis represents XylE reporter activity in milli-Units / min normalized to OD_{600}. Error bars represent the standard deviation of 3 technical replicates. **C)** In addition to the wildtype construct, three additional "scrambled" constructs were built for each of the *pmoC, pqqA,* and *ssrA* promoter predictions: one with the -35 hexamer replaced with the -35 anti-consensus sequence (AAAGGC), one with the -10 hexamer replaced with the -10 anti-consensus sequence (ACCCTC), and one with the middle 6 nucleotides in the spacer replaced with a random hexamer optimized for the greatest hamming distance relative to each of the *pmoC, pqqA,* and *ssrA* spacers (AGTAAG). **D)** Expression results for the scrambled sequence variants of *pmoC, pqqA,* and *ssrA.* XylE activity is measured as described in B).

Previous work characterizing housekeeping promoters in *E. coli* found that the -35 and -10 hexamers were the most important sequences for transcription initiation.[105] To verify that the -35 and -10 hexamers within our predictions carried the core transcription initiation signal, we created additional *xylE* reporter constructs with a section of the

predicted promoter's sequence scrambled (Figure 3.8.C). Specifically, we started with the wildtype promoter prediction and independently replaced either the -35 or the -10 hexamer with an anti-consensus sequence derived from the consensus motif, hypothesizing that these scrambled variants would disrupt transcriptional activity. As an additional control, we created a construct where we scrambled the middle six bases of the spacer. The spacer sequence may partially influence transcription initiation[106,107] although spacer sequence length, rather than composition, is believed to be of greater importance.[105] Thus, we hypothesized that constructs with scrambled spacers but intact -35 and -10 hexamers would have comparable transcriptional activity to the wildtype prediction.

We conducted scrambled sequence experiments for three top genes: *pmoC, pqqA*, and *ssrA*. We found that in all three cases, disrupting either the -35 or -10 hexamer was sufficient to disrupt XylE activity, however disrupting a hexamer in the middle of the spacer showed no discernable difference from wildtype XylE activity (Figure 3.8.D).

Several predicted promoters did not show strong XylE activity in the reporter assay, even though sequencing verified the promoter and *xylE* sequences in the constructs. It is possible that other factors were important in obtaining high transcript levels of the genes involved with these promoters. These results demonstrate the importance of experimental verification, once predicted promoters are obtained.

3.5. Pipeline validation with model organisms

After experimentally validating promoter predictions in a non-model organism, we additionally assessed the performance of the entire computational pipeline on previously published expression data for two well-studied bacteria using two transcriptomics methodologies: *Bacillus subtilis* (gram positive, microarray [108]) and *E. coli* (gram negative, RNA-seq [63]). For *B. subtilis*, the framework found sets of highly expressed genes across 11 experimental conditions. The consensus sequence compiled from the promoter predictions consistently reflected the known consensus for the housekeeping sigma factor (sigma-A) across a wide range of top *n*% thresholds (Figure 3.9.B). For *E. coli*, the consensus compiled from top promoter predictions across 10 experimental conditions matched the sigma-70 motif in the -10 hexamer for *n*=8-10%, however the canonical -35 hexamer was not clearly reflected (Figure 3.9.A). In a previously published analysis of promoter information content comparing 8 species of bacteria, Latif *et al* [109] reported that *E. coli* promoters have the lowest overall motif information content (known promoters tended to have more variable sequences relative to the consensus) while other organisms, including *B. subtilis* and *Thermatoga maritima,* skewed very high in promoter information content (known promoters tended to have

very conserved sequences relative to the consensus). This insight is consistent with the information content trends seen for the *E. coli* and *B. subtilis* motifs compiled from our framework's promoter predictions, with *M. buryatense* motifs falling in between (Figure 3.9.C).

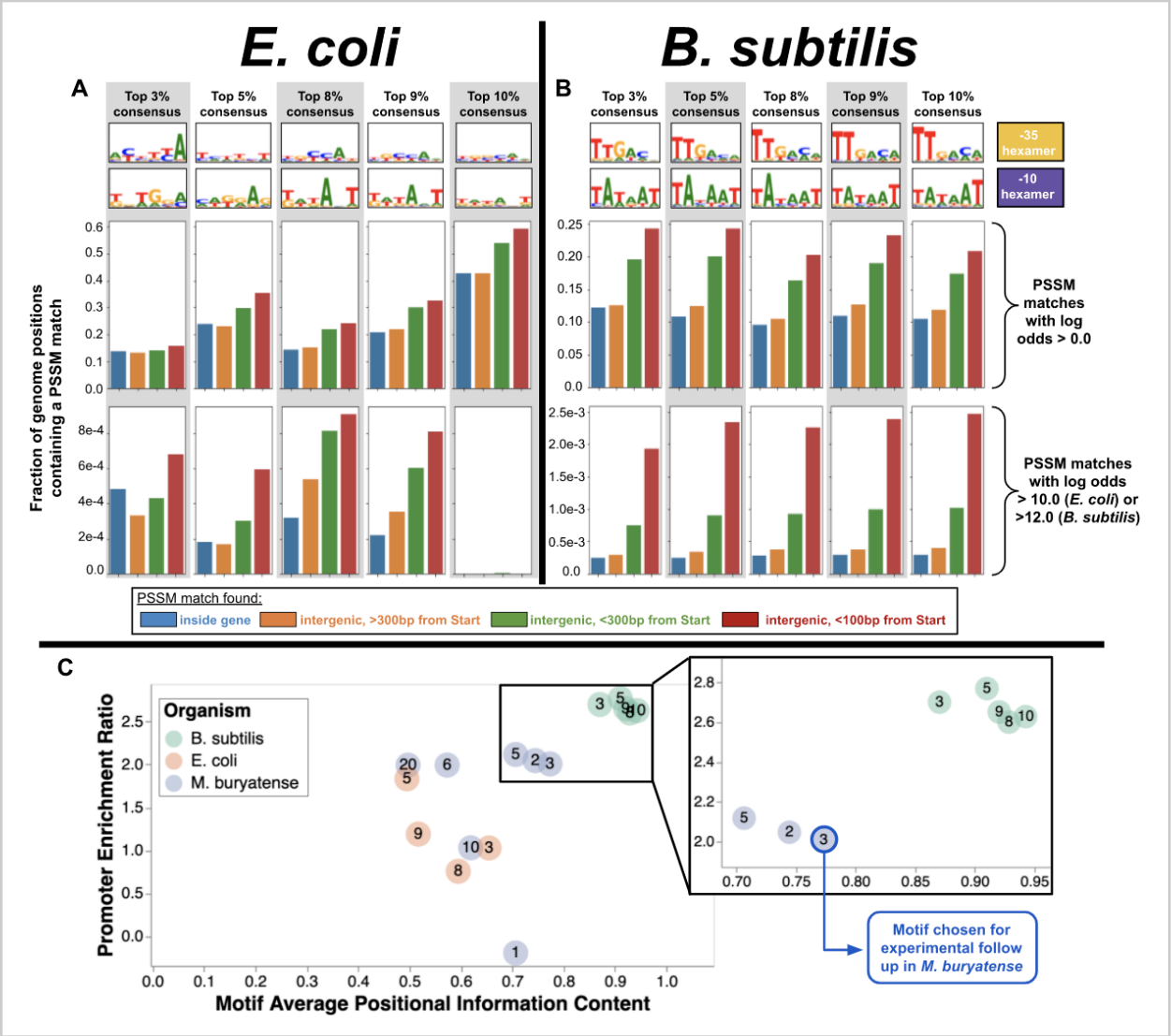


Figure 3.9. Validation of computational framework in model organisms. **A)** Consensus motif predictions for the top 3%, 5%, 8%, 9% and 10% sets of top expressed genes in *E. coli* and their relative frequencies in four genome locations, as in Figure 3.7. The -10 hexamer matches the canonical *E. coli* housekeeping motif for 8-10% but is not reflected clearly for the -35 hexamer. **B)** Same as A) but for *B. subtilis*. The canonical housekeeping promoter is clearly reflected for the -35 and -10 hexamers across all top n% gene sets tested. **C)** Visualization to compare the quality of predicted motifs for different gene sets across different organisms. Each point is a predicted motif from *M. buryatense, E. coli,* or *B. subtilis*. The numerical label on each point is the top *n*% gene set used to derive that motif. The x-axis represents the information content of each motif averaged across each of the 12 positions in the motif. The y-axis represents the $\log_2$ ratio of the frequency of motif matches <100bp from a gene start to the frequency the motif was found in intergenic regions (ratio of red bar to orange bar for each predicted motif in panels A and B).

It seems likely that our approach may not identify a precise promoter consensus in organisms with naturally heightened variability in their promoters. However, this framework should still be useful for identifying strong promoters as tools and is appropriate for non-model organisms with less underlying variability and no previously-known promoter tools. Since promoter variability may not be known before conducting this analysis, we provide an additional tool for displaying information content of identified motifs versus enrichment of the motif in promoter regions, as in Figure 3.9.C. Users may compare the framework-predicted consensus for a new organism with our results from *M. buryatense, B. subtilis,* and *E. coli* as a guide.

3.6. Summary of contributions

The computational workflow described in this chapter relies solely on standard whole genome and RNA-sequencing experimental data that are straight-forward and routine to collect for most prokaryotes. We applied our pipeline to the industrially promising methanotroph *M. buryatense* 5GB1 and report the following biological contributions: 1) a set of 25 constitutively, highly expressed genes in all growth conditions tested (publicly explorable via interactive visualization: erinhwilson.github.io/promoter-id-from-rnaseq); 2) a -35 and -10 consensus motif for constitutive strong expression in *M. buryatense*; and 3) six experimentally validated 30bp sequences that can be used to drive strong expression in this organism in a synthetic cassette.[110]

This effort resulted in a more thorough characterization of the relationship between promoter sequence and strength in *M. buryatense* and discovered several promoters not previously characterized. Not only do these findings contribute an expanded expression toolkit to improve our ability to effectively engineer *M. buryatense* for industrial biomolecule production processes, but the computational framework is open source and may be similarly applied to tease apart key pieces of regulatory grammars in other non-model organisms with limited experimental data. Ultimately, this framework should help grow the potential of metabolic engineering platforms to more flexibly pursue alternative organisms and develop a bioeconomy based on sustainably-sourced materials.

Chapter 4. Identification of iModulons in *M. buryatense*

4.1. Gene regulatory networks

With a suite of strong, constitutive promoters, initial headway can be made to install and test the feasibility of heterologous pathways in a new microbial host. However, achieving maximum efficiency in molecule production is not synonymous with inducing maximum possible expression of pathway genes. Excessive overexpression of non-native pathways can actually be detrimental, as host organisms may be unable to handle the additional metabolic burden of producing target molecules at extremely high fluxes or become redox imbalanced.[111]

As scientists move towards more fine-tuned optimization of pathway configurations for their organism, they must additionally have access to nuanced expression tools that can more carefully balance metabolic fluxes. Decoupling an organism's growth phase from its molecule production phase has been shown to benefit productivity, enabling more efficient metabolic engineering systems.[112,113] Cells can dynamically swap between these states during product generation if the transcription initiation of the growth and production modules are properly regulated.

A common tool to better control the timing of a culture's growth versus production phases is a genetic switch: a functional regulatory element that can alter gene expression in response to a change in the environment.[114] Examples of environmental changes include the addition of a small molecule to the growth medium, or simply a change in temperature or light.[114,115] Though cheaper materials will help processes remain economical, anything that can be controlled experimentally and trigger a signal cascade that activates or represses transcription, translation, or protein function could be a genetic switch option.

Regulatory interventions that repress expression at transcription initiation are ideal for conserving cellular resources, as synthesizing proteins costs significant energy.[116] Promoters that are inducible or repressible with relatively low-cost changes to the growth parameters are of particular interest for their potential to serve as metabolic switches.[114,117] In addition to the core promoter, the dynamics of transcription initiation are largely controlled by networks of transcription factors (TFs) that recognize and bind to particular DNA motifs, known as transcription factor binding sites (TFBS).[29,118] TFBS exist throughout the genome but are often concentrated in promoter regions; they tend to be short (6-12 bases) and can be arranged in varying combinations.[24,30] Cells

interpret these TFBS patterns and use them to perform logical operations to determine which genes need to be activated or repressed in response to the current environmental conditions. Co-regulated genes often share similar TFBS in their promoter regions and thus categorizing genes into modules based on their transcriptional response patterns is a key first step for uncovering regulatory motifs with the potential to be developed as expression tools.[119]

4.2. Methods for discovering groups of co-regulated genes

Given the intricate relationship between transcription initiation and an organism's genetic grammar, combining biological expression data with promoter sequence data to gain insights into gene regulatory mechanisms is a well-trodden path. After all, the ability to reliably predict phenotype from genotype is perhaps one of the most sought after endeavors in modern molecular biology.[120]

A seminal platform that combines expression data with motif finding is cMonkey[121] and its extension cMonkey2.[122] Designed to infer global gene regulatory networks, cMonkey is able to integrate microarray data, common motif patterns in upstream DNA sequences, and functional annotation networks to discover biclusters: clusters based on both genes and subsets of experimental conditions. The goal of the cMonkey work was primarily to discover functional biological modules and their regulatory networks, in particular for the archaeon *Halobacterium* NRC-1. The authors' approach to incorporate motif and function data on top of expression data helped to constrain this vastly under-constrained clustering problem and allowed them to detect biclusters that recapitulated known biology as well as make novel predictions. While cMonkey's framing of how to integrate expression data with motif finding is quite useful, an approach that does not rely on external annotations (e.g., metabolic pathway associations, protein-protein interactions) could be more generalizable to non-model organisms as these data are not always available when exploring organism hosts on the frontier of our knowledge.

When TFBS sites are not yet known for an organism, new TFBS can be characterized via experimental methods, such as ChIP-seq,[123] as well as with de-novo motif-finding techniques, such as PWM scanning,[124] expectation maximization,[84,125,126] markov models,[100,127] and many others.[102] Another approach that has seen recent success is independent component analysis (ICA): an unsupervised matrix decomposition technique.[128] Originally developed to parse independent signal sources from mixed audio recordings, ICA has previously been applied to microarray gene expression data, and more recently, to RNA-seq data in a wide array of organisms.[129,130] Notably, ICA has

been shown to outperform various types of biclustering methods for gene module detection tasks in model organism and synthetic datasets.[131]

As a brief summary, the intuition for using ICA in transcriptomic contexts is the conceptual similarity of overlapping “speaker signals” present in audio recordings taken throughout a noisy room to the idea of mixed “transcription factor signals” that contribute to expression levels in RNA-seq measurements in a variety of different growth conditions. Specifically, ICA decomposes an expression matrix into two parts: 1) the Module matrix where each gene is assigned to one or more independently modulated groups, each of which is regulated by a distinct signal source (i.e. a transcription factor), and 2) the Activity matrix, an estimate of the overall expression signal strength of each independent component in each of the experimental conditions measured. Analyzing these separated matrices can provide biological insights about the transcriptional network structure of the organism by elucidating sets of genes that are influenced by the same expression signal. Furthermore, by estimating the conditions in which each signal source is most active, ICA can indicate the likely biological function of that signaling mechanism.[63]

A database of ICA analyses for detecting independent gene modules (iModulons) from gene expression matrices has been growing rapidly.[130] Named iModulonDB, the accessible open-source workflows available along with a sophisticated interactive user interface made this an attractive method for discovering regulated gene modules in a non-model organism like *M. buryatense*.

4.3. Characterization of iModulons in *M. buryatense*

The *M. buryatense* RNA-seq compendium is a promising dataset for ICA analysis. Though the number of experimental conditions explored is relatively smaller than several of the preliminary works in the iModulon database in model organisms like *E. coli* and *B. subtilis*, we anticipated that its diversity would still enable deeper characterization of several regulatory responses of interest.

In particular, a subset of conditions that measured RNA-seq over the course of a copper transition experiment was of high interest - our research group has previously observed a set of genes involved in a significant transcriptional repression response though the precise regulatory mechanism is not yet known.[132] As copper is a relatively cheap additive to the growth medium, it is a potential facilitator of a growth-to-production toggle switch for *M. buryatense* and deeper investigation of the signaling network controlling this gene module would be illuminating for genetic tool development.

Following the open-source protocols established by Sastry *et al.* (workflow repository: https://github.com/avsastry/modulome-workflow), we executed ICA analysis for iModulon detection. We prepared our RNA-seq data by calculating the log ratio of the TPM value for each gene in each sample relative to the average TPM of that gene in the baseline uMax (ideal maximal growth) condition. The workflow then runs PCA on the TPM matrix to reduce the dimensionality before running the ICA algorithm for 100 separate iterations. The ICA results from 100 runs were clustered using DBSCAN to detect robust modules that were detected in multiple ICA iterations. After repeating the robust module detection process for a range of principal components (i.e. the number of principal components used for ICA ranged from 20 to the maximum number of samples, increasing by steps of 20), the optimal number of principal components that results in the highest number of robust ICA-derived modules is chosen for a final ICA run. The final ICA run produces the Module and Activity matrices, after which a set of predicted iModulons are available for curation and investigation.

After executing the recommended ICA workflow, 43 iModulons were identified from the *M. buryatense* RNA-seq compendium (Table 4.1). If annotations such as GO terms of KEGG pathways are available, the workflow additionally enables enrichment analysis to aid in iModulon function characterization.

iModulon ID	Label	# of Genes
0	unclear	91
1	possible nutrient limitation transport	41
2	driven by single-gene (SG_1)	62
3	growth rate and transport related; energy systems down	178
4	copper repression	9
5	growth rate (translation)	223
6	SOS response	99
7	unclear	8
8	copper repression	20
9	driven by single-gene (SG_2)	1
10	membrane biogenesis	142
11	lanthanum repression	14
12	nutrient stress/chemotaxis	45
13	unclear	5
14	unclear	41
15	possible response to low methane	258

16	unclear	22
17	unclear	67
18	driven by single-gene (SG_3)	1
19	unclear	16
20	nitrogen restriction	62
21	partially growth rate-related	38
22	driven by single-gene (SG_4)	1
23	partially stress-related	88
24	driven by single-gene (SG_5)	1
25	stress response partial	18
26	unclear	5
27	driven by single-gene (SG_6)	1
28	growth rate subset	30
29	O2 limitation	39
30	nutrient stress	19
31	nitrogen limitation (Nif)	53
32	driven by single-gene (SG_7)	2
33	driven by single-gene (SG_8)	1
34	iron restriction	12
35	unclear	18
36	unclear	4
37	metal stress	143
38	nutrient stress partial	53
39	driven by single-gene (SG_9)	3
40	unclear	18
41	driven by single-gene (SG_10)	1
42	unclear	3

Table 4.1. Characterization of iModulons discovered in *M. buryatense.* Rows highlighted in green are worth following up experimentally.

Over half of the genes in *M. buryatense* were not assigned to any iModulon, possibly because those genes are not significantly regulated or they are not regulated in conditions covered in the dataset (Figure 4.1.A). While this result was concerning at first, the *E. coli* and *B. subtilis* ICA analyses from Sastry *et al.*[63] and Rychel *et al.*[64] showed similar trends. In our analysis, over 1,000 genes were assigned to 1 or 2 iModulons, and a few genes were assigned to as many as 8 or 9. The average size

across the iModulons was 46.9 genes, though most were smaller than 20 and a handful were greater than 100 (Figure 4.1.B).

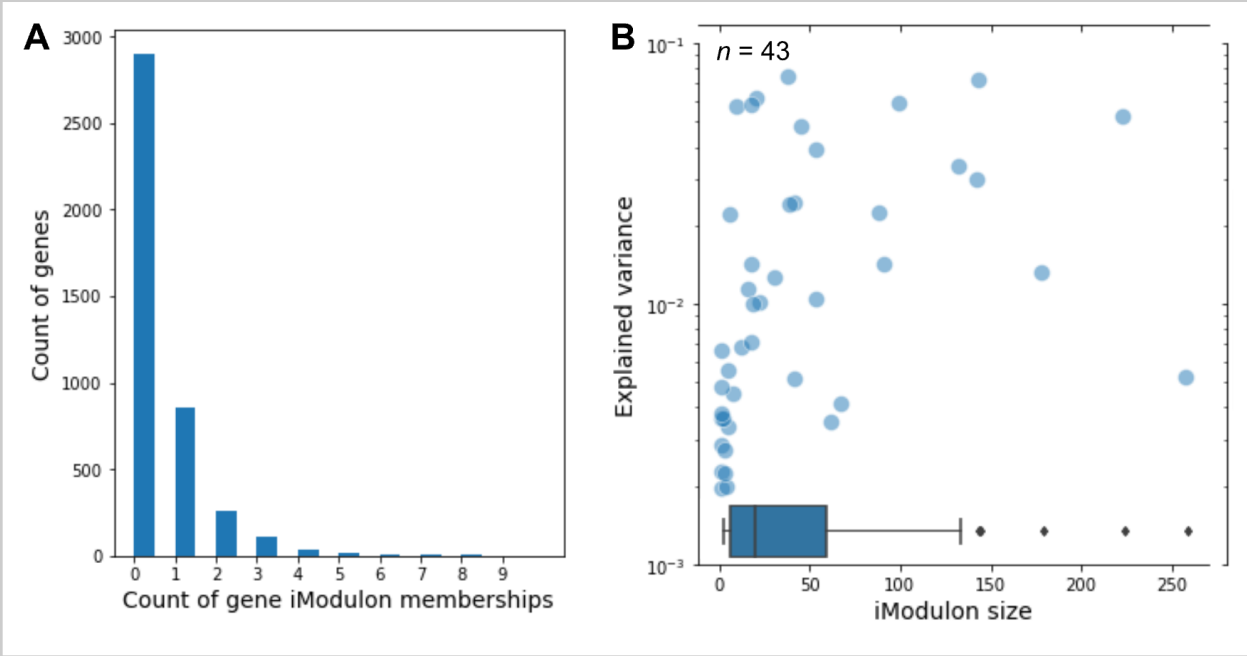


Figure 4.1. Summary of iModulon statistics in *M. buryatense*. **A)** Distribution of the number of iModulons to which individual genes were assigned. **B)** Scatter plot of iModulon size versus the amount of ICA variance explained by the gene members of the iModulon (*n* = 43 iModulons).

Several iModulons that recapitulate known systems in *M. buryatense* were immediately apparent. iModulon 4 clearly shows the sMMO gene cluster: the soluble methane monooxygenase that converts methane to methanol in the absence of copper (the pMMO enzyme is preferred when copper is present). This iModulon contains only 9 genes but explains a large portion of the variance from the ICA analysis (Figure 4.2.A) and these genes are all colocated on the genome (Figure 4.2.B). The activity of iModulon 4 is, as expected, highest in the “No Copper” condition and tapers off as copper increases (Figure 4.2.C).

Interestingly, there is a second iModulon that is active when copper is absent: iModulon 8. Comprised of 20 genes, this module explains a similar level of ICA variance but is primarily located on a different area of the genome (Figure 4.2.D-F). iModulon 8 includes the copper binding protein *corA*, seven hypothetical proteins, several putative secretion proteins, and an RNA polymerase sigma factor EQU24_RS01900 that is much earlier on the genome. One member of iModulon 8, EQU24_RS19520, has been computationally annotated as a secretion protein and is notably the overall highest expressed gene in the No Copper condition, beyond even the highly expressed pMMO gene cluster. Its exact function is unknown.

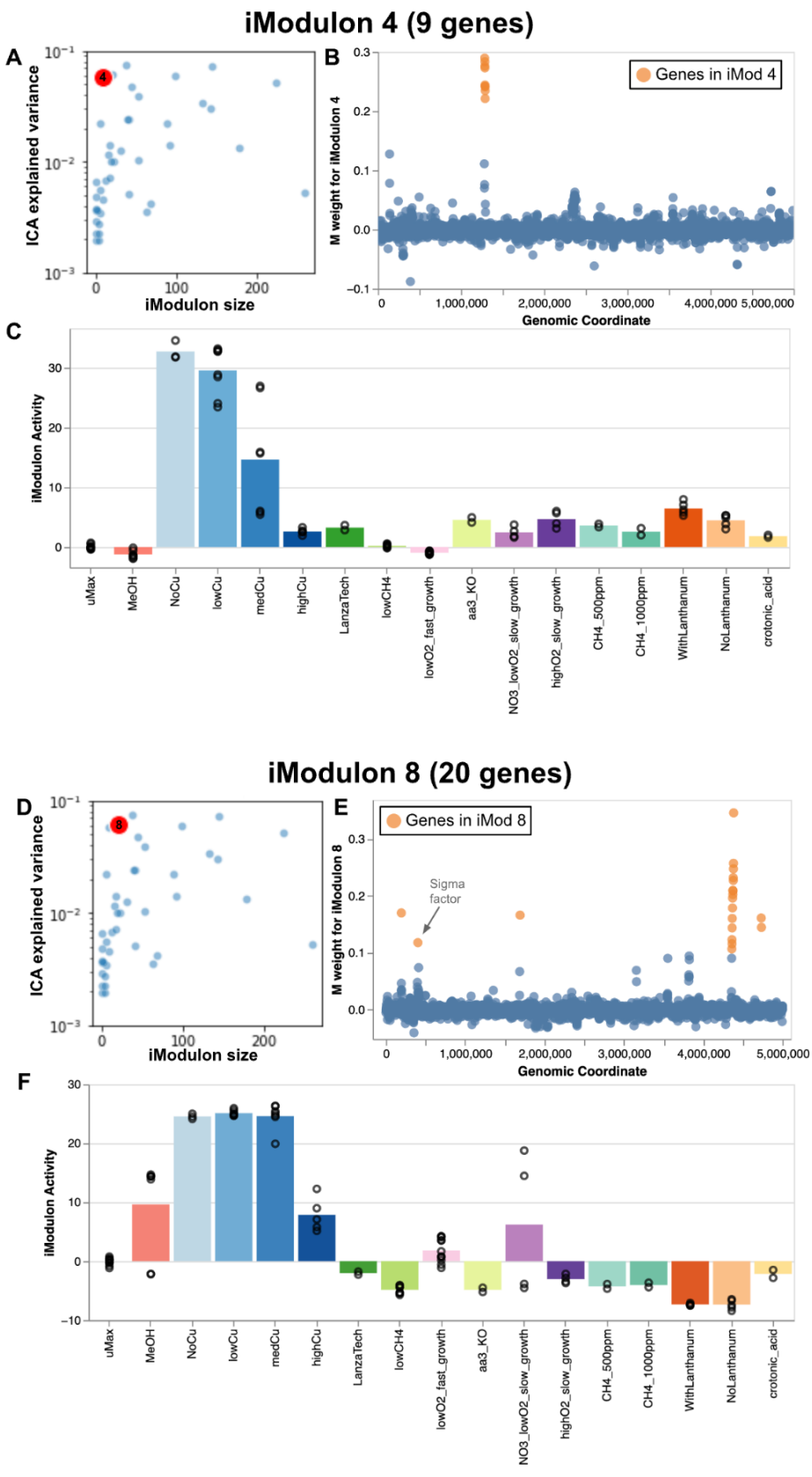


Figure 4.2. Copper repressible iModulons. **A/D)** Indicator of module size versus variance explained for iModulon 4 and iModulon 8, respectively. **B/E)** Module (M) matrix weight for each gene's contribution to the overall iModulon signal versus its position on the *M. buryatense* genome. **C/F)** iModulon Activity (A) vector across in all RNA-seq samples (black dots) averaged within each experimental condition group (bars).

Two other iModulons represent well known regulons in *M. buryatense*: iModulon 11, which contains lanthanum-repressible genes, and iModulon 31, which contains nitrogen fixation (Nif) pathway genes. Lanthanum has been identified as a repressor of the *mxaF* gene cluster for methanol dehydrogenase. When lanthanum is present, *M. buryatense* switches from the *mxaF* system to an alternative methanol dehydrogenase, encoded by *xoxF*. This switch can be observed in the stark drop in activity of iModulon 11 in the "With Lanthanum" condition relative to the "No Lanthanum" condition. The iModulon is also dramatically reduced in the crotonic acid condition. These samples were collected in unpublished work where it is suspected that strains accumulated a mutation that dysregulated the mxaF pathway. Notably, *xoxF* has the largest negative weight in this iModulon, though not quite enough to pass the threshold for being automatically called in this iModulon (Figure 4.3.A-C). On the other hand, iModulon 31 is quite a bit larger, containing a large Nif-gene cluster. Expectedly, this module is relatively active in the nitrogen-limited conditions "NO3_lowO2_slow_growth" and "aa3_KO" as the cells attempt to bring in more nitrogen resources (Figure 4.3.D-F).

Two other notable iModulons are 12 and 34, which are less well understood but appear to be nutrient stress and iron uptake-related, respectively. iModulon 12 contains 45 genes, many of which are chemotaxis proteins, suggesting that the cells were upregulating systems to move and search out nutrients when resources were low (Figure 4.4.A-C). iModulon 34 seems primarily driven by a response observed in two replicates from the lowO2_fast_growth fermenter runs and contains several iron uptake or iron transporter genes (Figure 4.4.D-F). It is likely these two fermenter runs became iron-stressed.

Most iModulons detected have been assigned preliminary functions based on activity profiles across the various growth conditions, gene members, and enriched GO term and KEGG pathway annotations. A handful of iModulons appear to be noise: the overall activity of the iModulon was largely driven by a single gene or there were many genes scattered along the genome that were included in an iModulon but the activity boundary between genes that were included versus excluded was not distinct. Noisy and single-gene iModulons also exist in the *E. coli* and *B. subtilis* analyses and are thought to be some combination of mixed or broad signals or a technical artifact. Such iModulons found in *M. buryatense* are not likely worth further investigation. A description of our current understanding of each *M. buryatense* iModulon is available in Table 4.1.

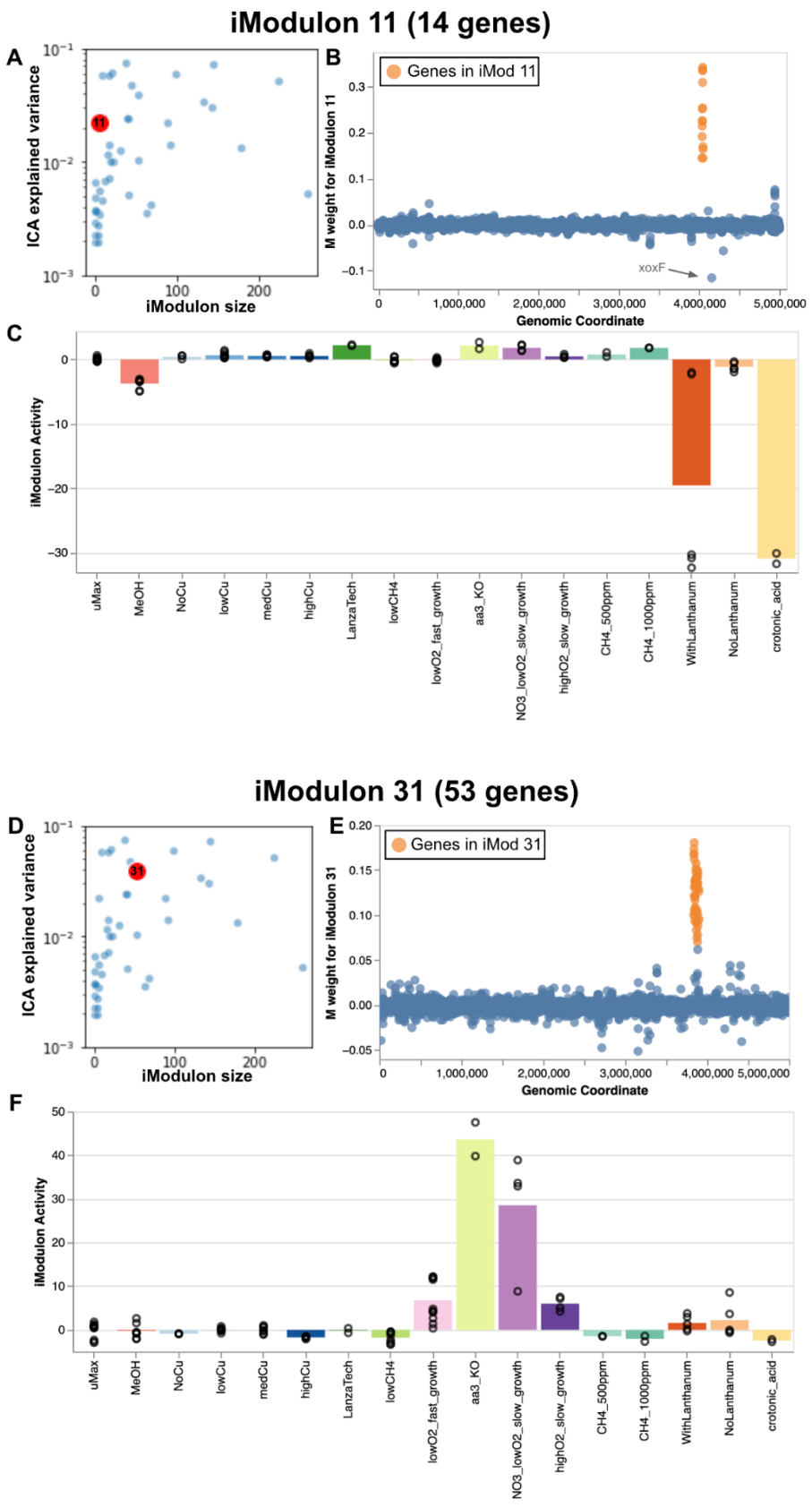


Figure 4.3. Lanthanum repressible (iModulon 11) and Nif-cluster (iModulon 31) iModulons.
A-F) Same as Figure 4.2 but for iModulon 11 and iModulon 31.

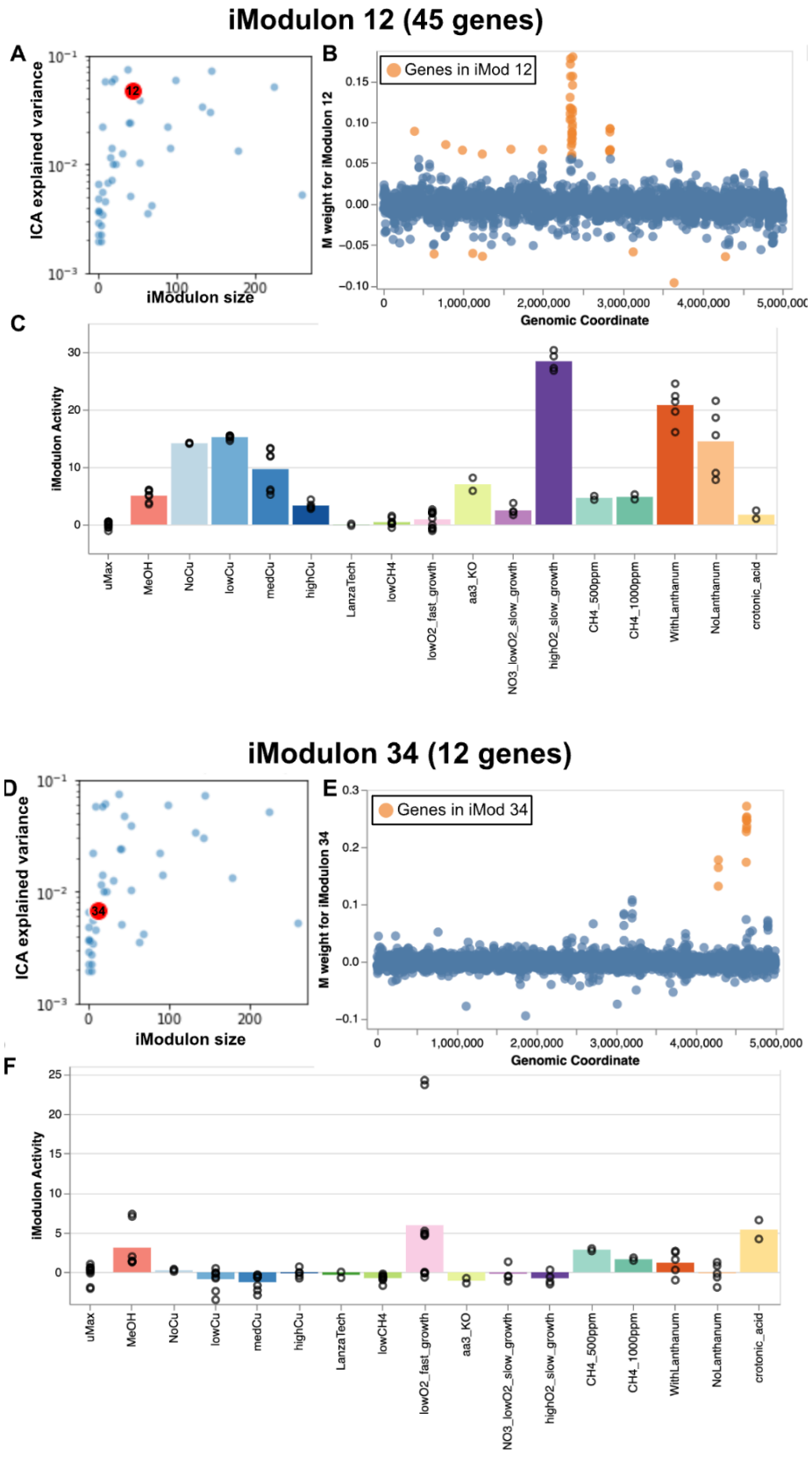


Figure 4.4. Nutrient limited (iModulon 12) and iron-uptake (iModulon 34) iModulons.
A-F) Same as Figure 4.2 but for iModulon 12 and iModulon 34.

4.4. Future directions for *M. buryatense* iModulon investigation

iModulon characterization was significantly aided by the inspection of iModulon activity across experimental conditions, gene expression information, gene set membership, and additional GO annotations. To enhance this curation effort, we took extra steps to provide a suite of interactive visualizations. The iModulonDB is an excellent visual resource and we adapted computational protocols to customize the interface into a series of dashboards hosted on a local server within our research group. Importantly, this interface enabled drilling down into specific genes and specific fermentor runs with hover tooltips, gene ID, and gene product search functionality, and clickable data points and tables that transport the user to another view with more specific information. These extra visualization tools were quite valuable in our curation efforts.

While the analysis described in this chapter is exclusively computational, it raised several hypotheses worth exploring experimentally. For example, previous efforts to knock out EQU24_RS19520 (the extremely highly expressed gene in the “No Copper” condition) have been unsuccessful. However genes within the nearby cluster as well as the sigma factor also identified in iModulon 8 are promising mutation or knockdown targets. Further investigation could improve our understanding of potential components in the *M. buryatense* copper-uptake or related secretion system that is distinct enough from the well characterized sMMO response to be in a separate iModulon.

Furthermore, iModulon 8 as well as the nutrient-limited iModulon 12 and the iron-limited iModulon 34 each contain a handful of hypothetical proteins. Experimental validation of the influence of these genes on *M. buryatense* fitness in the growth conditions in which each iModulon was most active could help further elucidate the function of a set of putative or unknown genes in this organism.

More broadly, this preliminary suite of gene modules can serve as a foundation for further development of genetic tools that control expression responses in specific growth conditions. Searching for shared TFBS motifs within promoter regions of iModulon members may elucidate specific sequence patterns that influence expression, and in fact, the iModulonDB workflow contains a section for applying the MEME-suite[126] of motif detection tools. Our past efforts applying MEME tools to *M. buryatense* promoter regions were not fruitful, however we were excited by the possibility of using iModulon groups as predictive labels in deep learning approaches detailed in Chapter 5.

Overall, by leveraging unsupervised machine learning techniques on a unique compendium of *M. buryatense* RNA-seq data, this work contributed to a broader characterization of diverse transcriptional responses in this non-model organism. While several experimental mutation targets have already surfaced as a result of exploring the

data through interactive visualizations, there remains much to be learned. This current estimate of independently modulated gene groups can be expanded and further tuned as additional RNA-seq data are collected, especially in new growth regimes. In the next chapter, we proceed with deep learning approaches to discover influential elements in the *M. buryatense* genetic grammar. As part of this effort, we incorporate the iModulon labels discovered here as prediction targets and explore the potential of additional regulatory structure to enhance model learning.

Chapter 5. Probing the limits of deep learning for genomics in data-limited regimes

5.1. Learning sequence-to-function relationships

Genetic grammars are largely composed of short DNA sequence patterns that are scattered throughout the genome. These patterns range from nearby motifs that recruit transcription factors to bind and activate or repress transcription, to distant motifs that influence DNA conformational changes to similarly promote or reduce transcription.[20] Building models to predict these “sequence-to-function” relationships is an ongoing effort across computational biology, especially in medical contexts where elucidating variants that influence disease-causing transcriptional dysregulation has major therapeutic potential.[133,134] But similarly for synthetic biology, identifying key motifs as well as understanding the influence of their specific combinations and arrangements on transcription activation is essential to efficiently engineering organisms to execute novel gene expression programs for biomolecule production.

Regulatory motifs can be discovered and characterized via experimental methods, such as ChIP-seq[123] and ATAC-seq.[135] Additionally, many computational motif-finding techniques have been developed, such as PWM scanning,[124] expectation maximization,[84,125,126] markov models,[100,127] and many others.[102] A newer suite of approaches has recently gained traction with the rise of machine learning. In particular, deep learning - a subset of machine learning that focuses on multi-layered networks capable of learning non-linear relationships between inputs - has emerged as a promising tool for detecting regulatory motifs and other efforts in synthetic biology.[136–138]

Deep learning excels at automatically detecting patterns in unstructured sequences of data. For example in image classification tasks, without explicitly instructing a model to find eyes, noses, and mouths when identifying pictures of humans, models have been shown to learn these features on their own.[139] This approach is similarly well-suited for many biological tasks, especially in regulatory genomics.[136,140,141] One of the first deep learning models developed for genomics was DeepBind,[78] a model trained to predict experimentally determined protein binding scores for a dataset of variable-length sequences using a convolutional neural network (CNN). CNNs are a type of deep learning architecture that use learnable sliding filters, which can discover patterns even when the location of the pattern within the input is unknown. Alipanahi *et al.* [78] were able to reliably predict DNA and RNA protein binding specificities while other prominent work such as DeepSea[79] and Basset[80] successfully trained CNNs to recognize TFBS sites in ChIP-seq data and predict accessible genome regions from DNase-seq data,

respectively. Recurrent neural networks (RNNs) with long short-term memory (LSTM) have also been explored and can improve performance on these tasks.[142,143] The number of publications using deep learning to make significant improvements in biological prediction tasks continues to expand to topics such as chromatin state, mRNA abundance, and protein design.[133,144,145]

Another successful tactic in learning sequence-to-function relationships has been training models on Massively Parallel Reporter Assay (MPRA) data. The combinatorial search space for DNA sequences is incredibly vast. Even for relatively short sequences of 50bp, the total possible sequences, 4^{50}, is more than the number of stars in the observable universe (~10^{23}).[146] The number of genes in the genome is but a tiny fraction of the search space, and so an alternative approach is to synthesize randomized sequences and experimentally measure their influence on gene expression. The key insight is that even though many sequences will be nonsense in the language of an organism's grammar, some will by chance contain relevant patterns or motifs, such as binding sites, that have a measurable influence on expression. Though testing all possible sequences and learning their functions is experimentally intractable, testing hundreds-of-thousands to millions of random sequences is already a much wider scope than using native genes alone. Several notable works train deep learning models on MPRA data measuring regulatory effects within the 5'UTR,[72,73] 3'UTR,[75,147] and promoters,[76,77] and can predict expression levels directly from DNA sequences with high success. However, these experimental protocols are quite specialized and would require intensive development to be adapted for many non-model organisms.

Many of the initial and high profile deep learning models for biology focused on human genomes, and in particular, the potential to elucidate rules of non-coding DNA sequences with relevance to various medical conditions. However, work focusing on microbial gene regulation has grown as well. A number of models, such as BaccPP,[148] bTSSfinder,[149] and CNNProm[150] aim to classify sequences by their likelihood of being a promoter while others estimate promoter strength directly from its sequence.[151,152] The tool iPSW(PseDNC-DL)[153] attempts both to classify sequences as promoters and subsequently estimate its strength by using both a CNN architecture combined with pseudo dinucleotide composition information (a vector of length 16 containing 2-mer frequencies).

Notably, all of these methods benefit from access to pre-curated databases of known promoters and their strengths. Most frequently used is RegulonDB,[89] a compilation of both computationally predicted and experimentally validated promoter sequences in *E. coli*. While RegulonDB is a valuable resource as a benchmark dataset, the tendency to over-engineer tools to fit a particular benchmark can be risky in machine learning.

Showing modest improvement over competitors on a benchmark task is less meaningful if it causes a loss in generalizability to other non-benchmark tasks.[154]

Expanding modeling efforts to interrogate organisms without such well-curated databases is difficult. To do so, we must frame modeling tasks such that useful patterns can still be learned even in the absence of ground truth annotations. For example, if promoter sequences have not yet been formally mapped or identified in an organism, evaluating the accuracy of True or False predictions for sequences that are candidate promoters will likely be unreliable. However if expression data exist where genes are measured to be differentially active, predicting the behavior of a candidate promoter, such as its tendency to increase or decrease expression, may be feasible instead.

5.2. Exploration of deep learning techniques for *M. buryatense*

5.2.1. Initial modeling approach

The diversity of successful prediction efforts that apply deep learning models for ‘omics tasks prompted us to explore their application for predicting *M. buryatense* transcriptional responses directly from the DNA sequence of its promoter regions. Especially since a curated database of promoters and transcription factor binding motifs is not available for this organism, we wished to determine whether convolutional models that can automatically detect patterns from the data could better enable their discovery.

Our previous attempts to discover regulatory patterns used the MEME-suite[126] of motif detection tools, however after lengthy troubleshooting, we were not able to find promising motif signatures that were both significant and widely prevalent within *M. buryatense* promoter regions. It is possible the MEME-suite tools had more difficulty with these sequences due to noise in our promoter region dataset. Promoter annotations typically are based off of the transcription start site, however since these have not been explicitly mapped in *M. buryatense*, we use 300bp sequence windows upstream of the translation start site (start codon). Alternatively, *M. buryatense* may have a sufficiently different pattern or structure to its grammar that is more difficult for the expectation maximization framework in the MEME tools to capture. Ultimately, we are unsure of the reason for the largely insignificant MEME results but instead chose another approach.

Deep learning approaches offer an opportunity to jointly model transcriptional responses across many growth conditions at once as opposed to searching for motifs in one co-regulated group of genes at a time, as is standard in MEME-suite analyses. Since the *M. buryatense* RNA-seq compendium measures RNA counts across 17 different

growth conditions, we were especially interested in exploring multi-task deep learning methods for further decoding regulatory patterns that influence transcription across multiple different conditions. Because TFBS motifs may have regulatory relevance across multiple conditions, the multi-task setup would allow models to share learned features across related prediction tasks.[79]

The project described in this chapter set out to discover novel regulated promoter motifs in *M. buryatense*, with the eventual goal of extending the method to other non-model organisms. A big question from the outset was whether the RNA-seq compendium contained enough signal for deep learning models to adequately learn the nuances of gene regulation. After all, deep learning models are hungry creatures and thrive in big data regimes. While empirically, model performance tends to decrease as training data decreases,[155] it was not clear if our dataset fell beneath a threshold that was insurmountable for deep learning methods, and thus we proceeded with our investigation.

Our initial approach tested various deep learning model architectures to predict gene expression responses across a variety of growth conditions directly from genes' upstream DNA sequences (Figure 5.1.). Subsequently, we planned to use feature attribution methods[156–159] to identify patterns that most influenced models predictions.

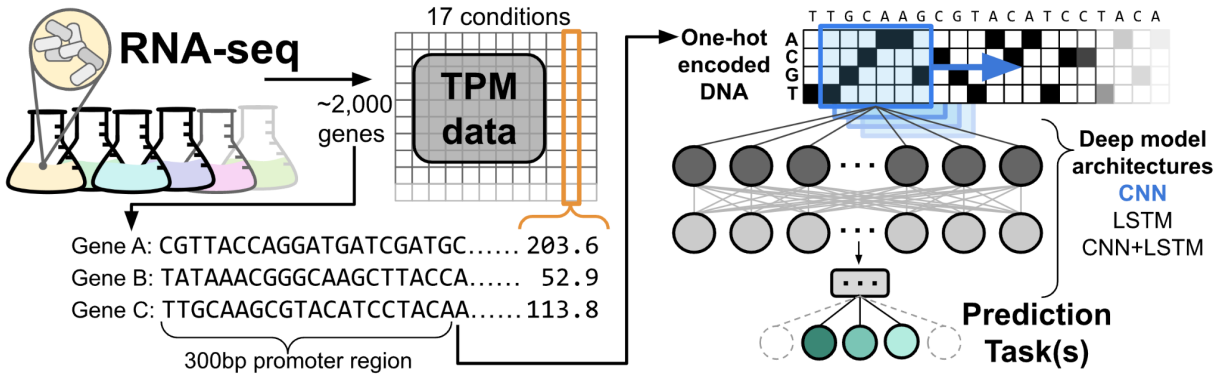


Figure 5.1. Overview of sequence-to-function learning approach. Transcript per million (TPM) values from the *M. buryatense* RNA-seq compendium were paired with the promoter region (300bp upstream of the start codon) for each gene. Promoter DNA was one-hot encoded and fed as inputs to deep learning models while the TPM values (or a transformation thereof) were used as the label (prediction target). Various deep model architectures, primarily hyperparameter variants of CNNs, LSTMs or CNN+LSTM combinations were tested.

To prepare our previously compiled RNA-seq dataset for deep learning tasks, we created two matrices **X** and **Y. X** is the input data matrix: 300bp DNA sequences immediately upstream of each of the 4,213 genes in *M. buryatense*. **Y** is the label

matrix: columns of measured TPM values for each gene in each RNA-seq experiment. As described in Chapter 3, RNA-seq experiments belonging to the same growth condition were averaged and genes likely to reside inside operons were excluded. After taking these into account, the transformed dataset consists of 2,204 gene upstream region examples (**X**) labeled with genes' average TPM values for 17 experimental conditions (**Y**). Notably, ~2,000 examples is a relatively small dataset compared to previous sequence-to-function learning efforts described in section 5.1, which typically trained on tens-of-thousands to millions of examples. Especially when divided into training, validation, and test sets at an 80% split ratio, learning complex genetics with ~1,700 examples is quite small indeed.

5.2.2. Results and challenges

A 17-way multi-task prediction framework is a fairly complex starting point (and difficult to debug should predictions fail). Therefore, we decided to start our implementation with a more focused question to ensure our approach can demonstrate robustness on simpler modeling tasks. This stepwise approach to start simple before building towards more complex goals helped us identify various challenges related to our dataset and make adjustments.

To reduce initial complexity, we updated our modeling target to be a subtask of the original goal: identify regulatory motifs that activate or repress gene expression in the presence or absence of copper. Copper is known to have a repressive effect on portions of *M. buryatense's* methane assimilation pathway[132] however the exact regulatory mechanism of this repression is not known. Given that a subset of genes have a measurable transcriptional response to variations in copper concentration, a TF binding mechanism is a plausible explanation. Additionally, copper is a relatively inexpensive medium component and thus a feasible material to use as a metabolic switch trigger for *M. buryatense*. Identifying a copper-responsive regulatory motif would be a highly useful outcome as we build out this organism's metabolic engineering toolkit.

Towards this more focused goal, we adjusted our data label matrix **Y** to only use conditions from two copper-related experiments: the "No Copper" growth regime and the "High Copper" growth regime. Specifically, for each gene, we calculated the log-ratio of its "High Copper" TPM value to its "No Copper" TPM value and replaced the 17-column **Y** matrix with this single, transformed column (referred to as the "log ratio copper expression value").

After analyzing our initial modeling efforts, we found that framing our prediction tasks as a regression on log ratio copper expression values was not feasible given the extremely narrow data distribution (Figure 5.2.A-B). Instead of learning to predict higher values for

genes relatively upregulated in “High Copper” and lower values for genes relatively upregulated in “No Copper,” every model configuration predicted the dataset mean (representative example, Figure 5.2.C). The range of log ratio copper values is not normally distributed as most genes in *M. buryatense* don’t respond to copper: only a small handful show a significant relative change between the “High Copper” and “No Copper” conditions. This indicated that the models were not actually learning to predict the variable expression of this small copper-responsive subset. Perhaps more accurately, it was never “worth it” for the models to risk predicting such outlying values (and take large penalties in the mean squared error (MSE) loss function): 95% of the genes were tightly distributed around the mean and thus MSE was best minimized by simply predicting the mean.

From a sequence design standpoint, it would have been useful to have a model that could predict continuous TPM values for a given input sequence and thus be minimized or maximized to tune a designed promoter up or down. However, we decided to move away from continuous-valued regression analyses that are drawn towards the overriding trends in data (in our case, not changing in response to copper) and instead reframe to a classification task in order to better capture genes in these outlying TPM ranges. With this adjustment, our genes were binned into one of three classes: “Up in Copper,” “Down in Copper,” or “No Change.” After analyzing genes ranked by TPM values coupled with expert biological knowledge of *M. buryatense*, we moved forward with 0.6 and -0.6 as thresholds: genes surpassing these thresholds for log ratio copper values were included in the Up and Down classes, respectively (Figure 5.2.D). After replacing the final linear regression layer of each model architecture with a classification layer, we next trained models to inspect genes’ 300bp upstream regions and predict the assigned copper class. For a three-class classification task, model performances are assessed using Cross Entropy Loss and overall prediction errors are reported with confusion matrices and class-balanced F1 scores (“Macro-F1”) to balance precision and recall across the three class labels.

In the realm of classification, our first challenge was class imbalance - a situation in which there is a large discrepancy between the number of examples depicting each class label, making it difficult to accurately classify instances from the minority groups.[160] With the -0.6 and 0.6 thresholds, the *M. buryatense* copper class distribution has 58 examples in the “Up” class, 145 in the “Down” class, and 1,956 in the “No Change” class, and thus highly imbalanced towards “No Change” with two minority groups. Unsurprisingly, our initial attempts at classification resulted in most models exclusively predicting the majority “No Change” class for every example seen (representative example, Figure 5.2.E,H). To mitigate class imbalance, we started by oversampling the minority classes (SMOTE)[161] and found that this helped models start to overfit to the

training data rather than solely predict the majority class, and thus learn nothing (representative example Figure 5.2.F,I). However, SMOTE still resulted in highly overfit models, as the predictions on the validation set remained poor (class balanced-F1 score = 0.32) (Figure 5.2.I) and the validation loss curves starkly diverged from the training loss (Figure 5.2.G).

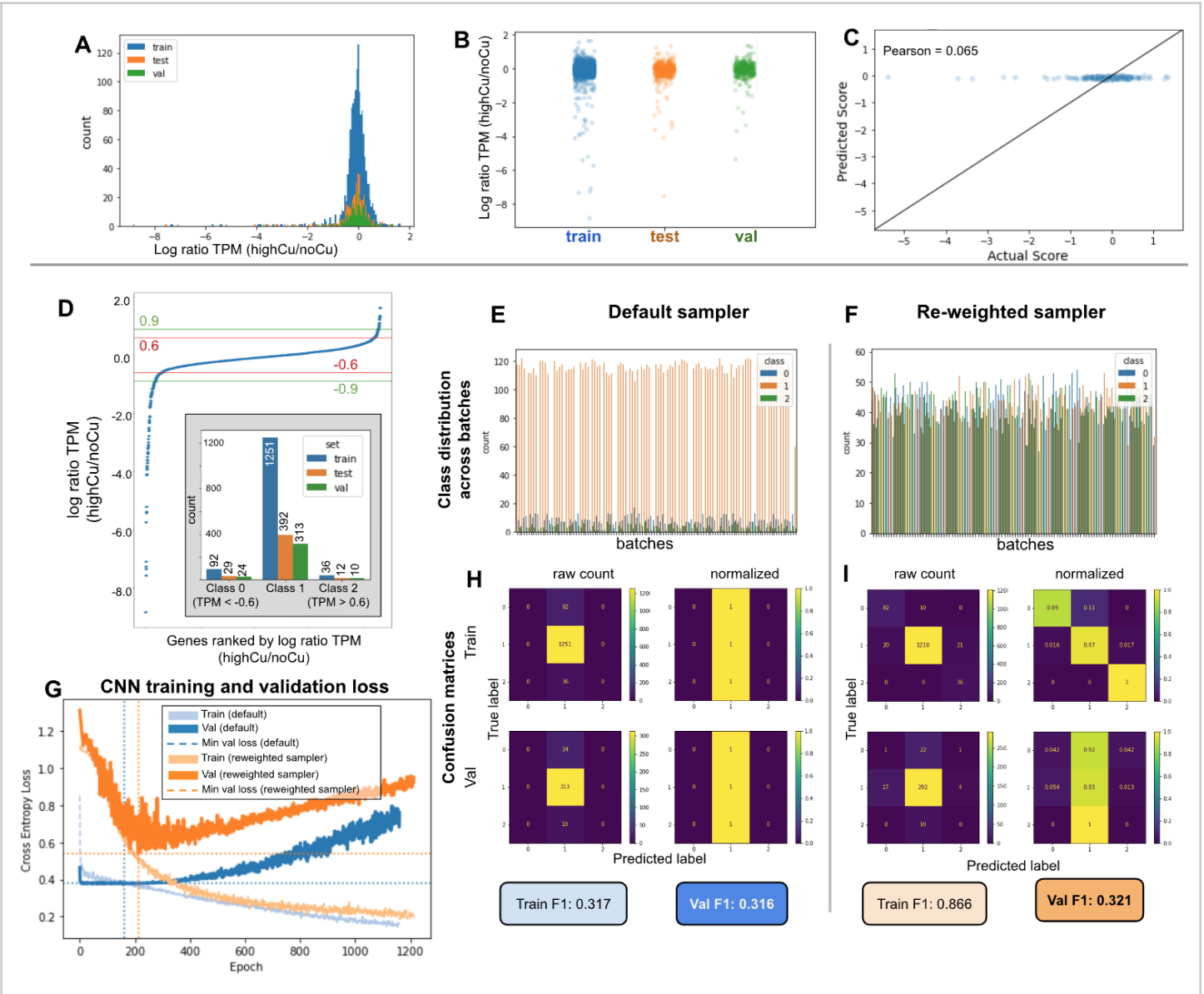


Figure 5.2. Initial deep learning results for *M. buryatense* copper prediction tasks. **A)** Histogram showing distribution of log ratio copper expression values (log_2 ratio of genes' "High Copper" TPM value to "No Copper" TPM value) across train, validation, and test splits. **B)** Strip plot showing same distribution, more clearly emphasizing the small handful of extremely negative ratio values. **C)** Representative example of a parity plot for a CNN model attempting a regression prediction: all predictions are simply the mean of the distribution. **D)** *M. buryatense* genes ranked by the log ratio copper expression values. At a threshold of +/- 0.6, genes were split into 3 classes (inset). **E)** Sampled class distribution across each training batch. **F)** Sampled class distribution across each training batch using SMOTE to rebalance the frequency minority classes are included. **G)** Training (light lines) and validation (dark lines) loss curves over training epochs for

default (blue) and SMOTE (orange) sampling strategies, measured by Cross Entropy Loss. The epoch during which the minimum validation loss was achieved is marked with a dotted cross. In this case, SMOTE did not alleviate overfitting behavior. **H)** Confusion matrices for train set predictions (top row) and validation set predictions (bottom row) using the default sampling strategy. Confusion matrix colors are visualized as raw counts (left column) and counts normalized by total class size (right column). Labels: Class 0 = “Down in Copper”; Class 1 = “No change in Copper”; “Class 2 = “Up in Copper.” Overall F1-scores are reported for train and validation predictions below. **I)** Same as H, but for the SMOTE sampling strategy.

We pursued other class imbalance mitigation techniques, including re-weighting the loss function to incur higher penalties for mis-predicting minority classes, and augmenting the dataset by creating *in-silico* mutational variants or extracting multiple promoter regions from a wider sliding window around the translation start site. Additionally, we incorporated dropout, early-stopping and gradient clipping into our training process, conducted broader hyperparameter searches across other model types and architectures, and adjusted our train/test split algorithm to ensure highly similar promoter sequences were sorted into the same division of the dataset. We considered if the multi-task framework could enforce further constraints on the prediction task and re-framed the prediction task in a few other ways. Initially, we returned to our original idea: a multi-task regression to predict log-normalized TPM values in all 17 conditions simultaneously. Additionally, we attempted a multi-label classification task to predict gene iModulon membership from the promoter region based on the iModulon groups identified in Chapter 4 (Figure 5.3).

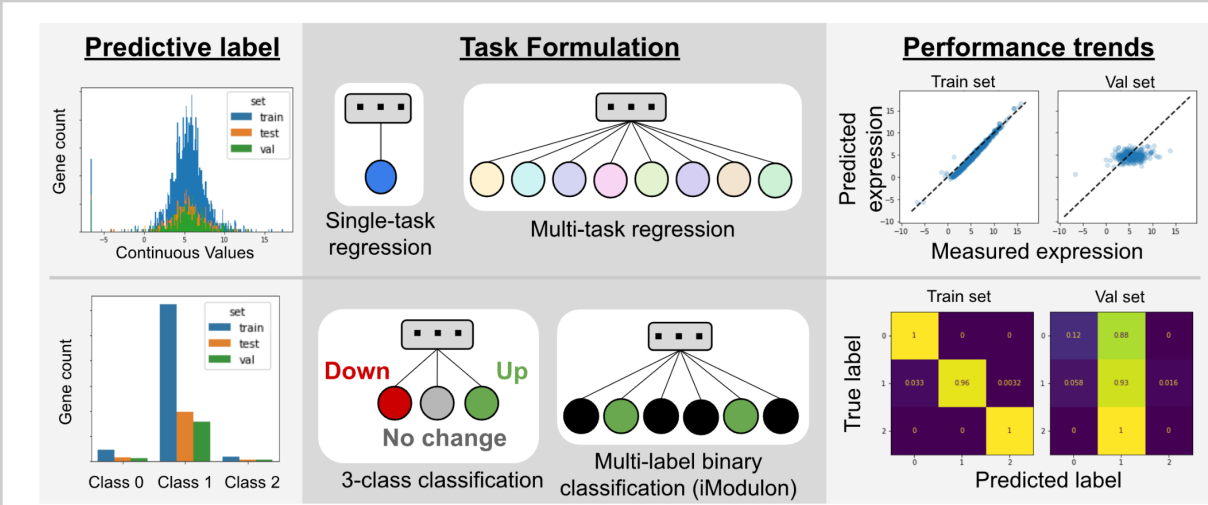


Figure 5.3. Schematic representation of performance trends observed for regression and classification task formulations. Continuous values (TPMs, log ratio of TPM relative to a baseline condition) were attempted in various single-task and multi-task formulations. Discrete class labels using log ratio TPM threshold to assign groups were tested, as well as multi-label binary classification tasks using iModulon labels. Despite various class imbalance, data augmentation, and overfitting mitigation strategies across numerous model types, performance trends primarily showed that models memorized the training data and did not generalize to the validation or test data (representative examples of overfitting parity plot and confusion matrix).

Unfortunately, none of the above tactics prevented the models from overfitting to the training data while learning patterns that generalized to the test data. For regression tasks, we primarily observed low Pearson correlation and R^2 scores when comparing the observed versus predicted values for each gene (Pearson score generally < 0.3). For the classification task, we instead observed low class-balanced F1-score (generally between 0.33 - 0.39). A representative example of a summary plot from a hyperparameter search that varied model architectures, learning rate, and data augmentation strategy to improve performance on the copper classification task is shown in Figure 5.4.

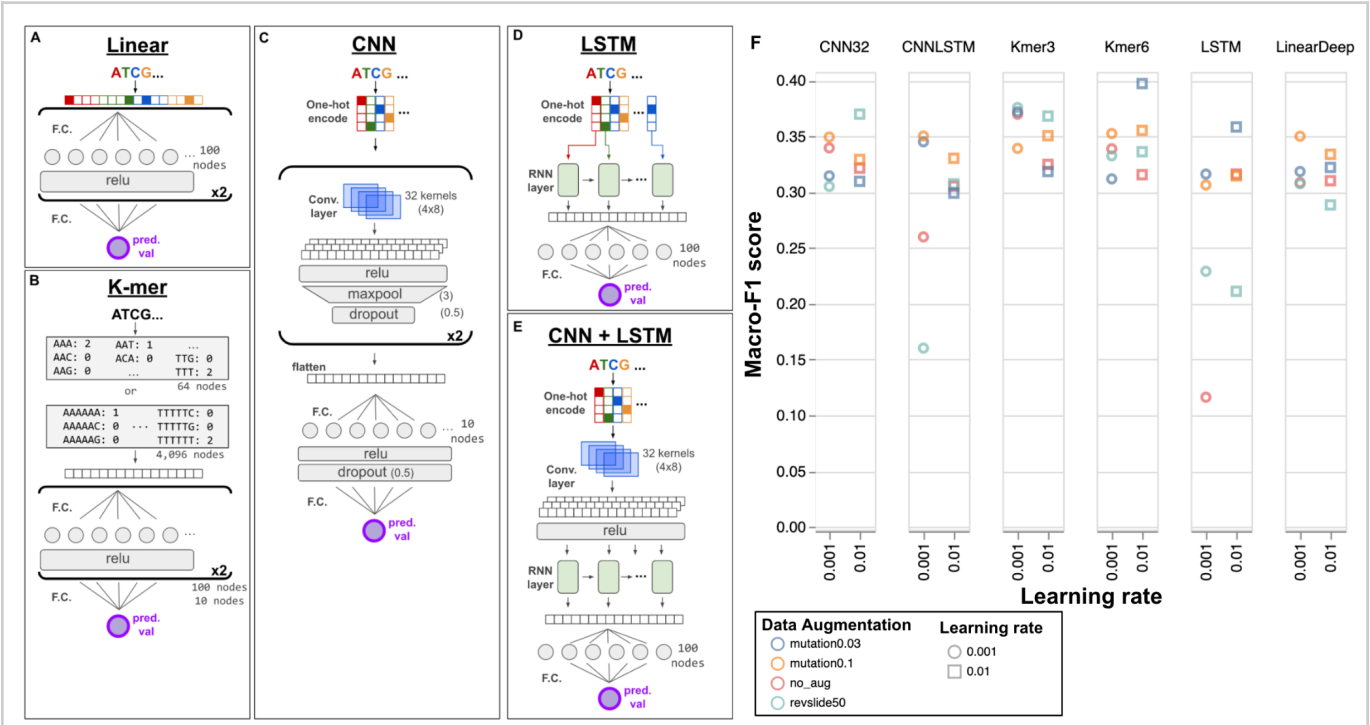


Figure 5.4. *M. buryatense* classification results from a hyperparameter search. **A-E)** Schematics of model architectures tested. **F)** Macro-F1-scores (average F1 over three classes to account for class imbalance) for models using a SMOTE re-weighted sampler to predict copper classes (defined in Figure 5.2.D). Data augmentation legend: “no_aug” = no data augmentation strategy used; “revslide50”: training data augmented by including reverse complemented sequences and sliding 300bp windows over promoter regions that have been extended by 100bp upstream and downstream, window stride of 50; “mutation0.03” = training data augmented by generating 10 copies of each training sequence and mutating single bases at a rate of 0.03; “mutation0.1” = training data augmented by generating 10 copies of each training sequence and mutating single bases at a rate of 0.1.

The difficulty of achieving reasonable model performance prompted us to reconsider some of our initial concerns about the dataset. Was the *M. buryatense* RNA-seq compendium too noisy? Are the number of gene examples in the genome too few to capture the complexity of its genetic grammar?

Given the ability of MPRA approaches to much more widely explore regulatory sequence space and provide extremely large datasets, we discussed the possibility of adapting a promoter MPRA to work in *M. buryatense* as a means of increasing the number of training examples. However, the relatively low transformation efficiency observed in past *M. buryatense* work would make such a high throughput experiment extremely difficult, suggesting this approach would have a very low likelihood of success. Furthermore, if deep learning methods are to be readily applicable to other non-model organisms with less extensive experimental protocols available, resorting to MPRA data would decrease the generalizability of the approach.

But this left us curious: if dataset size is the primary issue limiting model performance, how much data *would* be enough? To further investigate the limits of deep learning for predicting expression behavior in non-model microbes, we examined MPRA datasets in other organisms and a suite of synthetic prediction tasks where we evaluated model performance across varying degrees of data limitation.

5.3. Investigating performance limits of a systematically reduced MPRA dataset

As noted earlier, MPRA experiments have enabled deep learning models to learn biological features within random DNA sequences. The diversity of input sequences and the sheer number of examples to train on is well-suited to deep learning approaches and their power to learn influential motifs has been demonstrated to predict various types of regulatory phenomena.[72–76,147] While it is not surprising that more (high-quality) data generally leads to better predictive power, we were interested to stress the limits of data availability: we know these methods are useful with massive datasets, but at what point does their predictive power break down?

To explore this further, we used the MPRA from Cuperus *et al.*, which was used to predict expression enrichment in selective growth media from 50bp random sequences in an *S. cerevisiae* 5'UTR library.[72] We chose this dataset for a few reasons. First, it is large, but not gigantic: with ~500,000 sequence examples, the authors report moderately high correlation (R^2 score = 0.61) between measured and predicted values. Since our ultimate goal is to systematically reduce an MPRA down to the dataset size regime of the *M. buryatense* genome, this was a reasonable size from which to start reducing. Second, while *S. cerevisiae* is not a prokaryote like *M. buryatense* and the 5'UTR is not equivalent to the promoter region, it examines a key regulatory region in a model microbe commonly used in metabolic engineering and thus more similar to our desired context than human MPRA data. We considered a few *E. coli* datasets but the

ones we found used sequences that were genomically-derived instead of random, the predictive performance of the models described had much lower correlation values, or they were difficult to download and use.[162–165] The Cuperus *et al.* data was easily found and accessible on Github and had no major data wrangling hurdles that impeded its usability, which was a distinct advantage.

Our analysis approach was fairly straightforward: we trained models on the full training dataset, performing 5-fold cross-validation before evaluating each model on a held out test set. We then reduced the dataset by randomly downsampling sequences to make datasets that were 0.25, 0.025, and 0.005 fractional subsets of the original. Notably, the 0.005 sample setting reduced the Cuperus *et al.* dataset from 371,904 training sequences to 1,860, which is comparable to the *M. buryatense* training data size of 1,755. Additionally, we experimented with several model architectures, including a CNN, an LSTM, and a combined CNN+LSTM (Figure 5.5.A-C), as several papers published after Cuperus *et al.* showed that LSTMs can improve upon CNN performance.[142,166] In all cases, we repeated these training experiments on shuffled versions of the training sequences: shuffling a sequence preserves the overall GC balance, but it should destroy any biological motif signal due to the random rearrangement of bases. These models served as a null comparison to assess if models trained on sequences with true measurements were any better than expected by random chance.

To start, we trained a model on the full Cuperus *et al.* dataset and ensured that we could achieve the same predictive performance as in the publication. While we used a CNN architecture that was fairly similar in structure to the one described as optimal in their paper, it was actually quite a bit smaller: we used only 2 convolutional layers instead of 3, only 32 filters per convolutional layer instead of 128, filter kernels of width 8 instead of 13, and 10 fully connected nodes instead of 64. Despite its smaller size, this model performed just as well as the one published on the same test set (R^2 score = 0.62), indicating that the precise model architecture did not play a hugely important role.

We observed the expected trend of decreasing model performance with decreasing training data (Figure 5.5.D). The LSTM and CNN+LSTM models did not seem to achieve notably higher performance than just the CNN architecture at the full dataset size, though the LSTM performed nearly as well as the models trained on the full dataset as it did when trained on the reduction to 0.25. The 0.025 reduction setting saw all three architectures significantly drop in performance, with the LSTM being least effective in this degree of data-limitation. Lastly, at a reduction of 0.005 of the original dataset, models were barely better than those trained on randomly shuffled sequences, indicating a near total loss of predictive power (Figure 5.5.E).

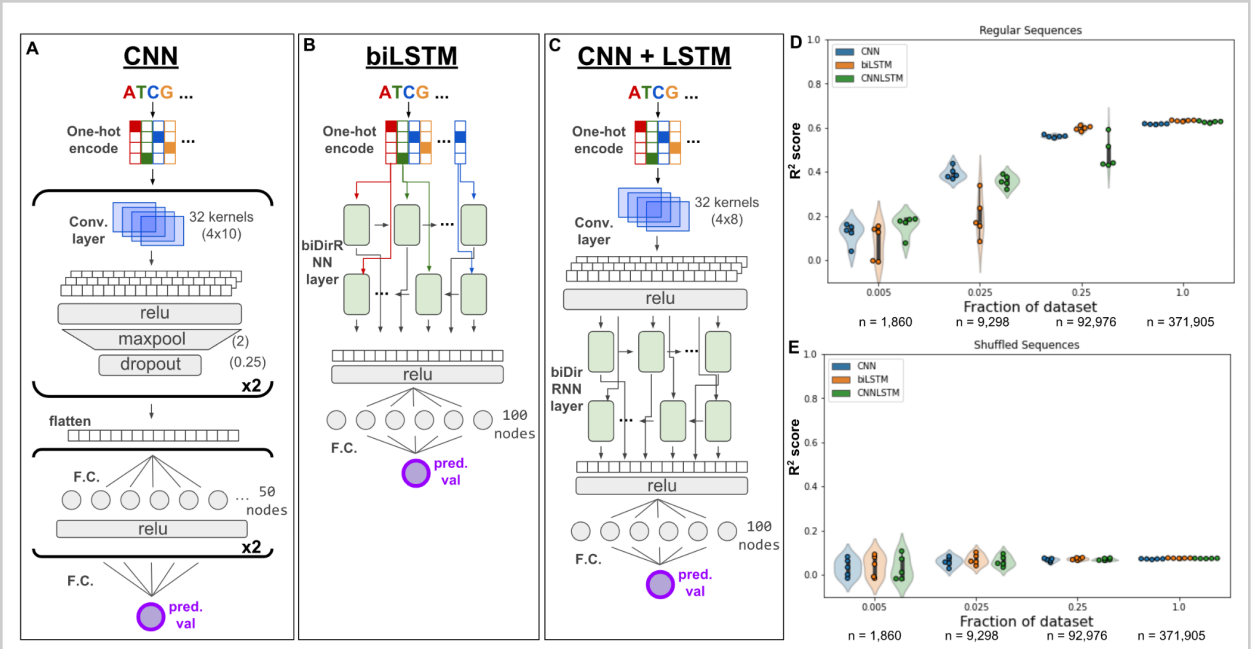


Figure 5.5. Deep learning model performance on a systematically reduced MPRA dataset. **A-C)** Model architectures tested in the MPRA reduction analysis. **D)** R^2 score of the experimentally measured versus model predicted values of test set sequences in the Cuperus *et al.* MPRA dataset. The x-axis groups models by training dataset size: the full training dataset (1.0) included ~370,000 sequences while the smallest fraction of the training set (0.005) included 1,860 sequences. Cross validation was used to train 5 different splits of the training set before evaluating on a held out test set defined by Cuperus *et al.* **E)** The same model architectures were trained on shuffled versions of the input sequences as a negative control.

This analysis provides useful insights as a point of comparison to the *M. buryatense* copper response prediction efforts described in Chapter 5.2. The Cuperus *et al.* MPRA dataset has many advantages over the *M. buryatense* promoter region dataset: it is about 250 times larger, has much more sequence diversity due to the randomness inherent in constructing the MPRA library, and the expression enrichment signal is bounded to 50bp sequences rather than 300bp. While eukaryotic organisms may have increased complexity, by most metrics, the Cuperus *et al.* dataset should contain a much clearer signal to noise ratio than the *M. buryatense* promoter region dataset. And yet, when the Cuperus dataset is reduced to the same size, it is nearly indistinguishable from random. This was our first major indicator that the size of our *M. buryatense* dataset was likely insufficient for a deep learning model to learn the complexity of its copper response.

According to the analysis in Cuperus *et al.*, a large driver of the expression enrichment signal had to do with the presence or absence of small open reading frames that happened to appear in the random 5'UTR library: these open reading frames were likely

inhibiting translation by confusing the ribosome about the true start codon. This biological mechanism is not likely a driving factor in the *M. buryatense* copper response and more likely due to an activating or repressing binding event. This prompted our next investigation: exploring dataset size limitations in the context of predicting a simulated binding event from random sequences with synthetic motifs.

A jupyter notebook tutorial walking through this MPRA reduction analysis is available on Github:
https://github.com/erinhwilson/mbur-sequence-learning/blob/main/cuperus_random_utr_prediction.ipynb

5.4. Evaluating models on a suite of synthetic prediction tasks across varying data-limited regimes

5.4.1. Defining a simple synthetic motif prediction task

We started by creating a synthetic example that simulated a common gene regulation scenario: a pair of motifs that activate or repress gene expression. We used a simple scoring function that created a small distribution of scores centered at 0, and gave sequences a +10 score bump if it contained motif-1 (activating motif) and a -10 drop if it contained motif-2 (repressing motif) (Figure 5.6.A).

To start simply, we used all the sequences in the *M. buryatense* promoter dataset (~2,000), sliced out the first 6 base pairs of each sequence, and scored this 6bp sequence by the synthetic function using “TAT” as the activating motif and “GCG” as the repressing motif. Any sequence with a total score of < -5 was assigned to Class 0, scores > 5 were assigned to Class 2, and everything else (without a motif) was assigned Class 1. Because this was such a simple task, we trained a CNN classifier with only a single convolutional layer, one fully-connected dense layer, and 32 3bp kernels to predict the class of each 6bp sequence.

Unsurprisingly, finding a 3-mer within a 6bp sequence was an extremely simple pattern for a CNN to learn: performance on the train and test set achieved perfect accuracy, and still achieved a 0.86 F1-score using a model with only 3 convolutional kernels instead of 32 (Figure 5.6.C, row i, iii). Conversely, when sequences were shuffled after they had been scored in order to disrupt the 3-mer patterns that had the biggest influence on the score, classification performance dropped dramatically (Figure 5.6.C, row ii). While the resulting high performance was expected due to the extreme simplicity

of this task, it provided reassurance that the modeling framework was not mis-configured.

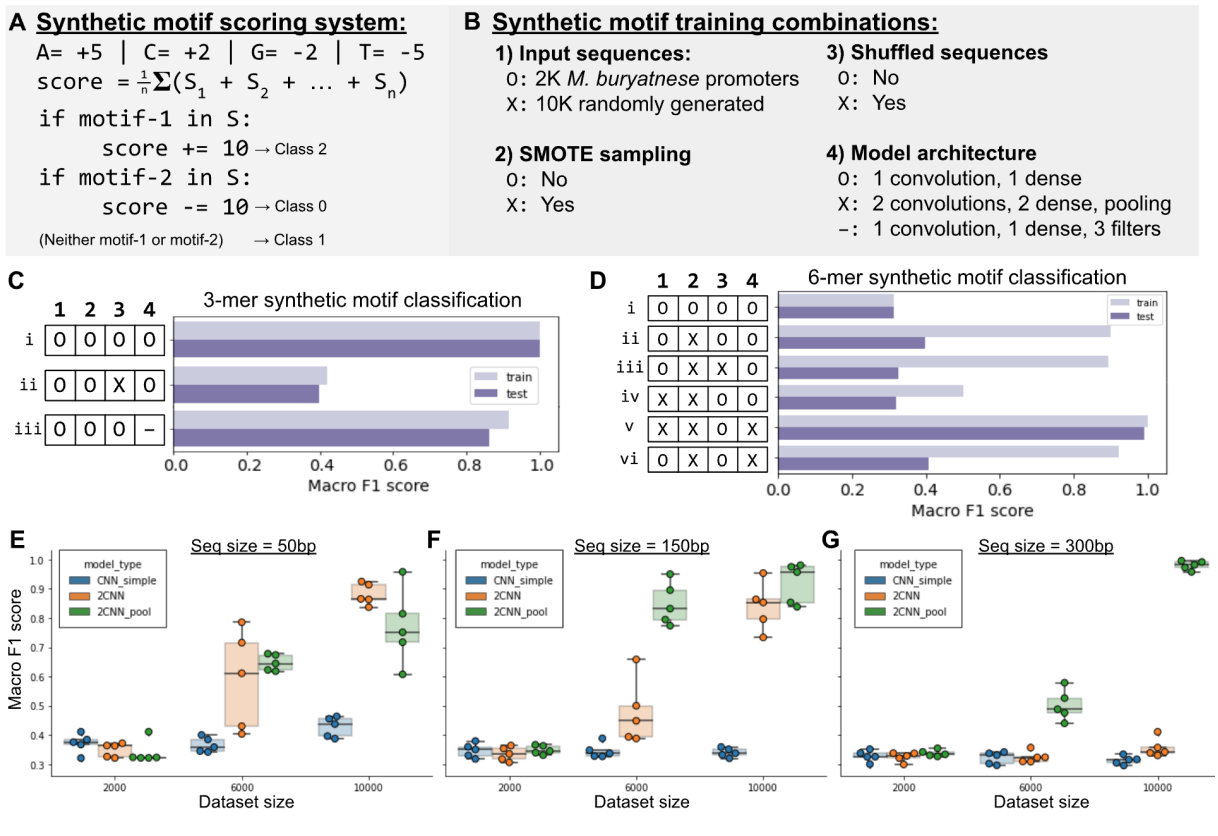


Figure 5.6. CNN performance on synthetic motif classification tasks. **A)** Sequence scoring system: A,C,G, and T are assigned arbitrarily point values, which are averaged across the whole sequence to create a score distribution centered at 0. Two k-mers are arbitrarily chosen and assigned as the activating or repressing motif, which gives a sequence a +10 or -10 score bump. Sequences were then assigned class 0 if the score is less than -5, class 2 if the score is greeted than 5, and class 1 if the score is between -5 and 5. **B)** Combination of training parameters used in initial exploration of synthetic classification task for 3-mer and 6-mer motifs. **C)** Classification results for synthetic 3-mer motifs in 6bp sequences using a very simple CNN with 1 convolutional layer and either 32 or 3 filters. Motif-1: "TAT"; Motif-2: "GCG". **D)** Classification results for synthetic 6-mer motifs in 300bp sequences using a few CNN architectures and input sequence datasets. Motif-1: "TATATA"; Motif-2: "GCGCGC". **E-G)** Summary of classification performance of 3 CNN architectures on randomly generated sequence datasets of varying size and sequence lengths (**E:** 50bp sequences, **F:** 150bp sequences, **G:** 300bp sequences). Legend: "CNN_simple": 1 convolutional layer, 1 dense layer; "2CNN": 2 convolutional layers, 2 dense layers, no pooling; "2CNN_pool": 2 convolutional layers, 2 dense layers, maxpool 3. Base architecture: filters=8, kernel size=6, dense nodes=10.

A tutorial demonstrating the results of this simple exploration is available here: https://github.com/erinhwilson/mbur-sequence-learning/blob/main/synthetic_DNAclassification_task1_simple.ipynb

5.4.2. Defining a more realistic synthetic motif prediction task

To more closely simulate the previous copper classification task with the *M. buryatense* data, we increased the complexity of the prediction task. As before, we started with the ~2,000 300bp upstream promoter regions from *M. buryatense*, but instead of scoring them by the log ratio of the measured TPM values in high versus no copper, we used the same synthetic scoring function as described in Section 5.4.1. The main change was to use 6-mer motifs instead of 3-mers, as biological binding sites are generally at least 6bp long,[167] and 300bp is a more realistically-sized sequence window to be searching through than 6bp as it is on the order of prokaryotic intergenic regions and read lengths from a sequencing run.[168] In this case, the activating motif was defined to be “TATATA” and the repressing motif was “GCGCGC.” To clarify, these motifs are not specific biologically relevant sequences - they were chosen arbitrarily. These motifs are being used as a demonstration of a signaling mechanism where we as the researchers know the ground truth (because we defined it) and are not guessing if the models are finding the “true” (score-influencing) signal or not.

The complexity for this new task has increased because the relative “signal density” within the search window has decreased from 0.5 (3bp/6bp) to 0.02 (6bp/300bp). We updated the basic model architecture to use 8 convolutional kernels of width 6, followed by one fully connected dense layer with 10 nodes (Figure 5.6.B). The rationale for this architecture is as follows: an overarching goal throughout these analyses was to keep models as simple as possible until it became necessary to increase the complexity. If we could find a boundary of the minimum compute resources necessary to accomplish a task of a given complexity, it would aid in recommending model architectures for other tasks given the amount of data and anticipated complexity of the signal being modeled.

With this basic model architecture, initial classification attempts failed in a similar fashion to the “copper log ratio expression value” models: the models exclusively predicted the majority class and did not learn any sequence patterns (Figure 5.6.D, row i). Class-balanced sampling did not recover performance nor did switching the dataset from the ~2,000 *M. buryatense* promoters to 10,000 randomly generated 300bp sequences: in these cases, the models were able to overfit to the training set, but did not generalize to the test set and were no better than when models were trained on shuffled sequences (Figure 5.6.D, rows ii-iv).

However, when we increased the complexity of the model by using two convolutional layers with a max pooling layer in between followed by two fully connected dense layers, the classification on the 10,000 randomly generated sequences was extremely successful, averaging 0.97 F1-score on the test set (Figure 5.6.D, row v). Interestingly, using this 2-convolutional layer model architecture with the smaller *M. buryatense*

promoter dataset did not work as well, achieving an F1-score of only 0.4 (Figure 5.6.D, row vi). This suggested to us that a combination of the larger dataset size *and* the more complex model factored into the performance improvements when the motif signal density was so much lower than 0.5.

To more extensively explore the relationship between model complexity, data set size, and signal density, we ran 5-fold cross validation for three different levels of each of these three factors:

- Model complexity:
    - Simple CNN: 1 convolutional layer, 1 dense layer
    - 2CNN: 2 convolutional layers, 2 dense layers, no pooling
    - 2CNN pool: 2 convolutional layers, 2 dense layers, max pooling
- Dataset size:
    - 2,000 random sequences
    - 6,000 random sequences
    - 10,000 random sequences
- Signal density:
    - 300bp sequence window
    - 150bp sequence window
    - 50bp sequence window

We found that for 300bp sequence windows, only the “2CNN pool” model worked in the largest 10,000 sequence dataset, though it had mediocre performance in the 6,000 sequence dataset (Figure 5.6.G). None of the models showed good performance for datasets containing only 2,000 sequences, which tracked with our observations of poor model performance with the equivalently small *M. buryatense* promoter dataset.

Interestingly at 150bp and 50bp where the signal density was higher, both the “2CNN pool” and the “2CNN” models had increased performance, though not quite as high as performance of “2CNN pool” models trained on 300bp sequences (Figure 5.6.E-F). The “Simple CNN” did not perform well in any dataset combination, though it was slightly elevated for the dataset with 10,000 examples of 50bp sequences. None of the model architectures tested could perform well with only 2,000 examples in the dataset at any sequence window size.

5.4.3. The influence of sequence length, motif prevalence, and data set size on model performance

The observation that the more complex model performed well in the 50bp and 150bp sequence regimes but not quite as well as the model in the 300bp regime might be

considered surprising: it would be reasonable to assume that shorter sequence windows should make the signals easier to find and therefore boost model performance to as good or greater than performance for 300bp sequences. However, this analysis did not control for the overall class balance across datasets: the chance is much smaller that a specific 6-mer motif sequence appears within a random 50bp sequence versus a random 300bp sequence. Therefore, the prevalence of “positive motif sequences” was much lower in the 50bp sequence regime, leaving a far lower total number of examples in the minority classes for the shorter sequence window datasets and creating more severe class imbalance.

To correct for this, we pursued an additional analysis where we controlled for the motif prevalence in each dataset. During random sequence generation, we ensured that the balance of motif-containing sequences included in the dataset were maintained at a specific percentage, which was now kept consistent across sequence lengths:

- Motif prevalence:
    - Minority classes each at 5% of dataset
    - Minority classes each at 1% of dataset

For this next analysis, we moved forward with only the “2CNN pool” model as it had shown the largest degree of performance variation, but we continued to vary the other three variables: dataset size, signal density (sequence length), and motif prevalence. Since the “2CNN pool” model performed nearly as well in largest 150bp dataset as the largest 300bp dataset, we investigated a narrower but more granular set of sequence window lengths: a ladder between 20bp and 150bp (signal density 0.3 to 0.04), taking 10bp steps within this interval.

Models trained on these synthetic datasets showed performance trends consistent with the hypothesis that if the motif prevalence and total training examples are kept equal, models perform better when the signal density is higher (shorter sequence windows) (Figure 5.7). This trend is demonstrated most clearly in Figure 5.7, panels A, D, E, and F. In panels B and C, both the motif prevalence and data set size appear to be sufficiently high for models to perform well at any signal density between 20bp and 150bp.

Intriguingly, there is high similarity between the trends in Figure 5.7.A and 5.7.F where the total number of motif-containing example sequences is the same: 5% of 2,000 sequences and 1% of 10,000 sequences both equal 100. This prompted us to consider if there exists a relationship between the “information richness” of the dataset and model performance that remains consistent across synthetic datasets. To calculate the information richness in each dataset, we multiplied the motif signal density (motif length

divided by sequence length), the motif prevalence (fraction of training data containing the motif), and the dataset size (Figure 5.8.A).

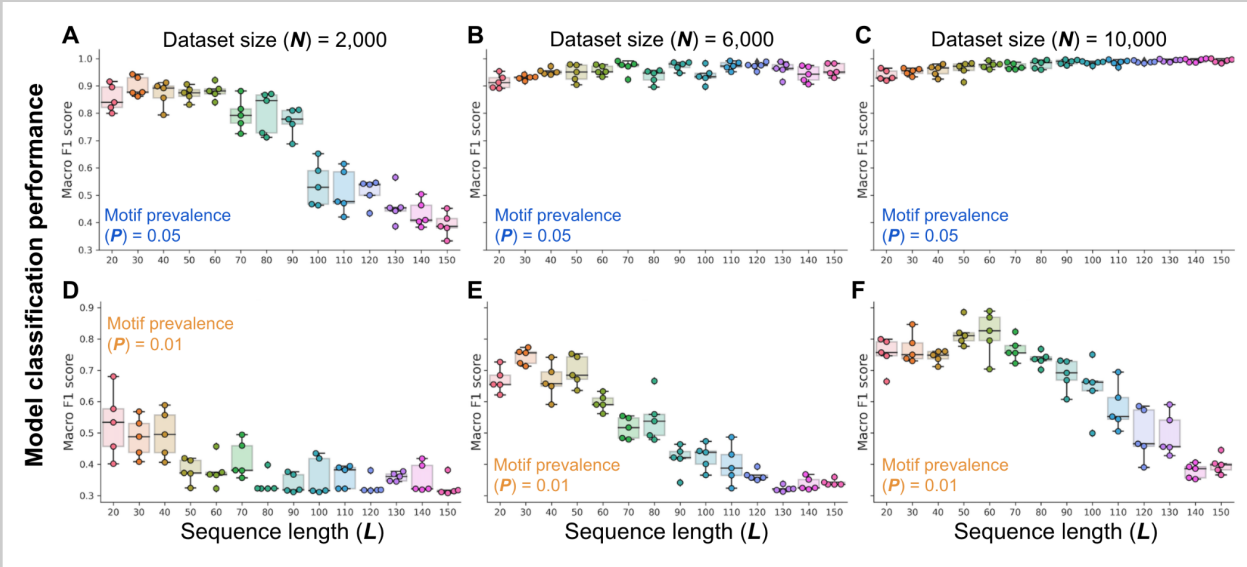


Figure 5.7. CNN performance on synthetic motif classification task with class imbalance controlled. Macro-F1 score was recorded for 5-folds of the training data for datasets of size 2,000, 6,000, and 10,000, using sequence lengths between 20bp and 150bp. **A-C)** The prevalence of motif-containing sequences for the minority classes is controlled at 5%. **D-F)** The prevalence of motif-containing sequences for the minority classes is controlled at 1%.

Plotting the information density for each dataset against the average model performance on that dataset yields a somewhat sigmoidal relationship (Figure 5.8.B). In particular, datasets with information density lower than ~6 seem to never perform much better than 0.6 F1-score in the classification task, regardless of the total dataset size. On the contrary, datasets with information density higher than 30 almost always perform better than 0.9 F1-score. As noted earlier in Figure 5.7.A,F, datasets of 2,000 sequences with 5% motif prevalence and datasets of 10,000 sequences with 1% motif prevalence contain the same total number of motif-containing example sequences. We saw these model performances within the range of sequence lengths explored track each other quite closely (Figure 5.8.C).

We do not expect the specific parameters defined by this curve to be a formal “law” - it is likely that the observed shape is additionally influenced by other factors, such as the precise model architecture, the motif signal lengths, and the complexity of the activation and/or repression mechanism. Further investigations are warranted to gain a more complete picture of this relationship. However, what *is* clear from this analysis is that for a task as biologically simple as one activating motif and one repressing motif, searching through 300bp sequences for ~6bp signals with only 2,000 examples is very difficult using a common CNN architecture that has been successful on other gene expression

prediction tasks, such as in DeepSea, Basset, and with MPRA data. Therefore, tasks that are more complex than a simple pair of activating and repressing motifs - which is most likely the case for true microbial genetic grammars - would be nearly impossible to predict in such a data-limited regime.

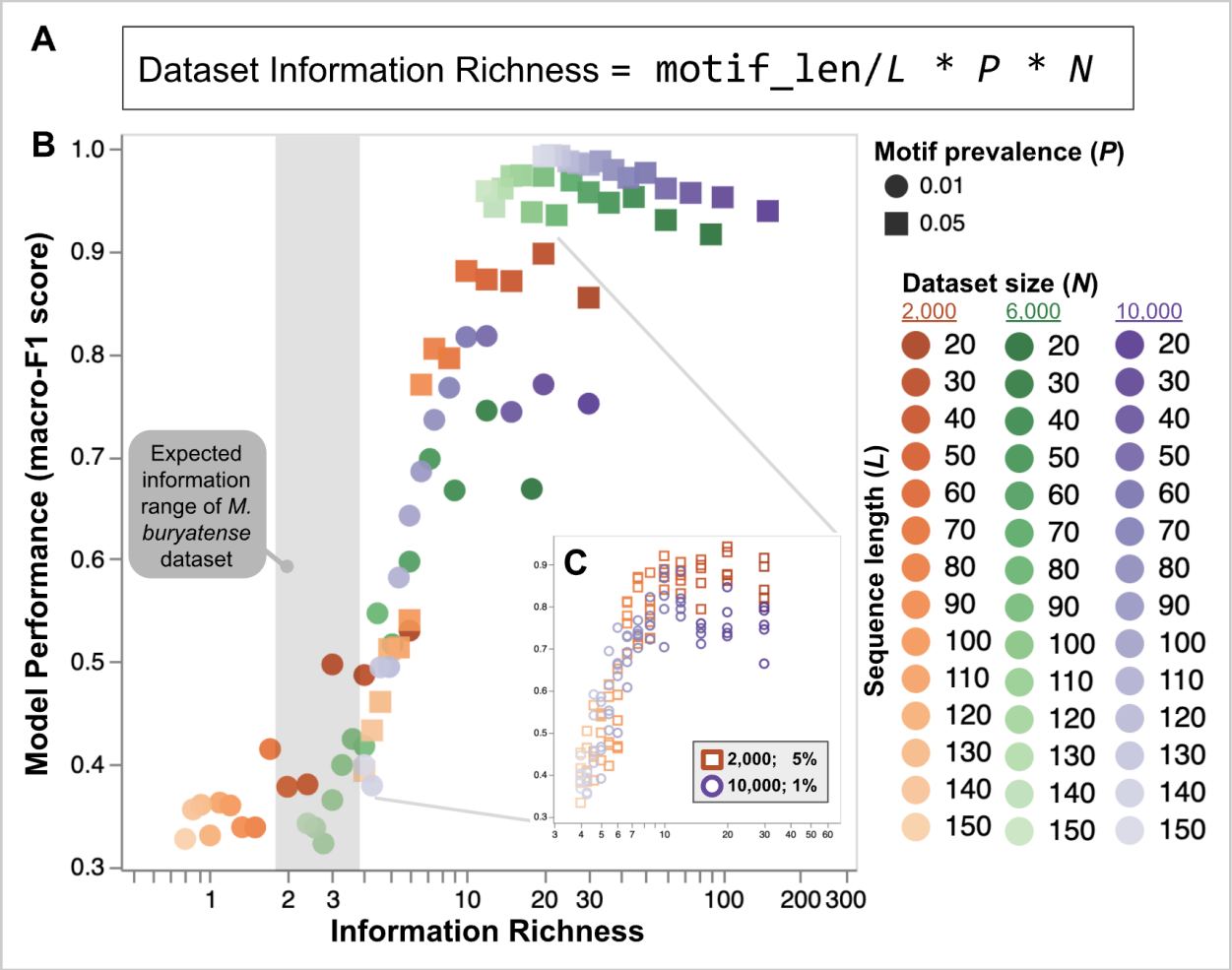


Figure 5.8. Model performance on datasets of varying information richness. **A)** Equation to calculate data information richness: "motif_len" = size of the signal motif (for synthetic classification task, motif_len=6 (motif-1: TATATA; motif-2: GCGCGC)); *L*: length of the inputs sequences, between 20bp and 150bp; *P*: prevalence of motif-containing example sequences in the dataset, either 5% or 1%; *N*: dataset size, either 2,000, 6,000, or 10,000. **B)** Relationship between dataset information richness and model classification performance (class-balance F1 score). Performance was averaged across 5-folds of the training data when evaluated on the held out test set. **C)** Subset of datasets where N = 2,000 and P = 0.05 versus datasets where N=10,000 and P=0.01. All 5 folds are plotted independently, instead of averaged, as in B).

5.4.4. Estimating the information richness within the *M. buryatense* promoter dataset for predicting gene expression response to copper

To tie this relationship back to our original copper prediction task for *M. buryatense*, we more precisely describe the promoter dataset with respect to these information richness levers. As stated, due to the limits of this organism's genome size and estimated operon membership, it contains only ~2,000 promoter region sequence examples. With transcription start site annotations available in other organisms such as *E. coli*, other CNN-based predictions efforts are able to use smaller sequence windows closer to 80-100bp. However, without such annotations available in *M. buryatense*, we use 300bp windows to ensure we do not fully miss the core promoter region, as the distance between the transcription and translation start sites can be variable within organisms.[169] If the transcriptional activation or repression observed within *M. buryatense* in response to copper is induced by a binding event, we can estimate a typical signal length between 6-12bp. Based on an upper and lower threshold cut off of 0.6 for the log ratio copper expression values (Figure 5.2.D), we assigned 145 sequences to the down-regulated class and 58 example sequences to the up-regulated class out of 2159 total, yielding an average motif prevalence of 0.046 between these two groups.

Plugging these values into the information richness equation in Figure 5.8.A, we can estimate that the information density with respect to copper response in the *M. buryatense* dataset is between 1.98 and 3.97 (Figure 5.8.B, shaded area). In the over-simplified synthetic task outlined above, datasets in similar regimes of information density typically achieved model performances between 0.3 - 0.4, which is similar performance to when models exclusively predict the majority class and ignore the minorities (macro-F1 score = 0.33 for three classes).

This suite of analyses help explain why the modeling efforts described in 5.2 were so unproductive. While there exist other possible factors that may be contributing to poor prediction performance – perhaps the copper binding signal is not located within the typical boundaries of prokaryotic promoter regions or perhaps this response is not actually regulated by a binding event – if there is a signal to find, is it not likely for standard CNN models to capture it with the *M. buryatense* dataset on its own given its limited information richness. We therefore conclude that incorporating additional data would be necessary to proceed with the prediction task, though it remains to be determined what additional data, if any, would be enough to improve predictive power.

A tutorial demonstrating the results of these synthetic motif explorations are available here:
https://github.com/erinhwilson/mbur-sequence-learning/blob/main/synthetic_DNAclassification_task2_harder.ipynb

5.5. Preliminary transfer learning results for *M. buryatense* copper prediction tasks

When the amount of training data for a particular task is too sparse to effectively train from scratch, one route for mitigating the large data requirements of deep neural networks is to use transfer learning: a process in which a model is trained for a different-but-related task with large amounts of data, and later fine-tuned on a specific task with less data available.[170] In these cases, the model for the secondary task is initialized using the weights learned during training for the first task, transferring knowledge by transferring learned features to help "warm start" training on the desired task. This approach has been shown to improve model performance in various biological contexts to predict gene expression behaviors across species,[171] cell types,[172] and transcription factors.[173]

To investigate if transfer learning could improve model performance for classifying genes into copper response groups by their promoter regions, we attempted three transfer learning scenarios. Each task used the same base model architecture:

- Convolutional layer with 32 filters, kernel size = 10, followed by ReLU
- Max pooling layer, pool size = 2, followed by Dropout of 0.25
- Convolutional layer with 32 filter, kernel size = 6, followed by ReLU
- Dropout of 0.2, followed by Flatten
- 2 Linear layers with 50 nodes each, followed by ReLU

The transfer tasks each had a final prediction layer with the number of nodes required by the specific task. After training on the transfer task, this final layer was replaced with a 3-class classification layer so that it could be fine-tuned for the *M. buryatense* copper prediction task after initializing the model with the learned weights from the transfer task.

To start, we collected performance data on "cold started" models which trained exclusively on the *M. buryatense* log-ratio copper expression values without any pre-training. We used 5-fold cross validation to assess a distribution of performances across different train/test splits of the ~2,000 genes and compared to models run on shuffled versions of the promoter regions. Shuffled sequences should preserve the overall GC balance in the dataset but destroy any biological motif signal, so this comparison served as a null model baseline from which to assess any performance improvements.

We observed that cold-started models (trained for copper response classification only) trained on wildtype sequences had slightly higher performance (class-balanced F1-score) than models trained on shuffled DNA (Figure 5.9.A). However the difference is not significant and models are primarily predicting the majority class (Figure 5.9.A, representative confusion matrix).

For our first transfer learning attempt, we started with a task based entirely on sequences from the *M. buryatense* genome: we used 300bp upstream regions from all genes as well as 300bp regions extracted from within coding sequences (genes that were longer than 300bp were split into multiple chunks for as many full 300bp intervals that could fit within the length of the open reading frame). This led to a slight class imbalance: upstream region examples were the minority class at about 25% of the dataset, though this is a much less extreme imbalance than ~5% (as was the case for minority classes in the copper response classification task). Models were trained to predict if sequences belonged to the upstream region set or the coding sequence set. The goal was to prime the model to learn features to differentiate the general grammar of promoters from that of protein coding regions, which could then be fine-tuned to learn features more specific to the copper response prediction task. We observed that this model architecture was well-suited for the transfer task, achieving an average of 0.92 F1-score on the test set measured during 5 fold cross validation (Figure 5.9.B). This was significantly better than the average 0.62 F1-score from training the same models on shuffled versions of the upstream regions and CDS sequences. While shuffled models were still mostly able to classify CDS sequences correctly, they had a much harder time with upstream sequences, more often incorrectly predicting them to be CDSs (Figure 5.9.B, representative confusion matrix).

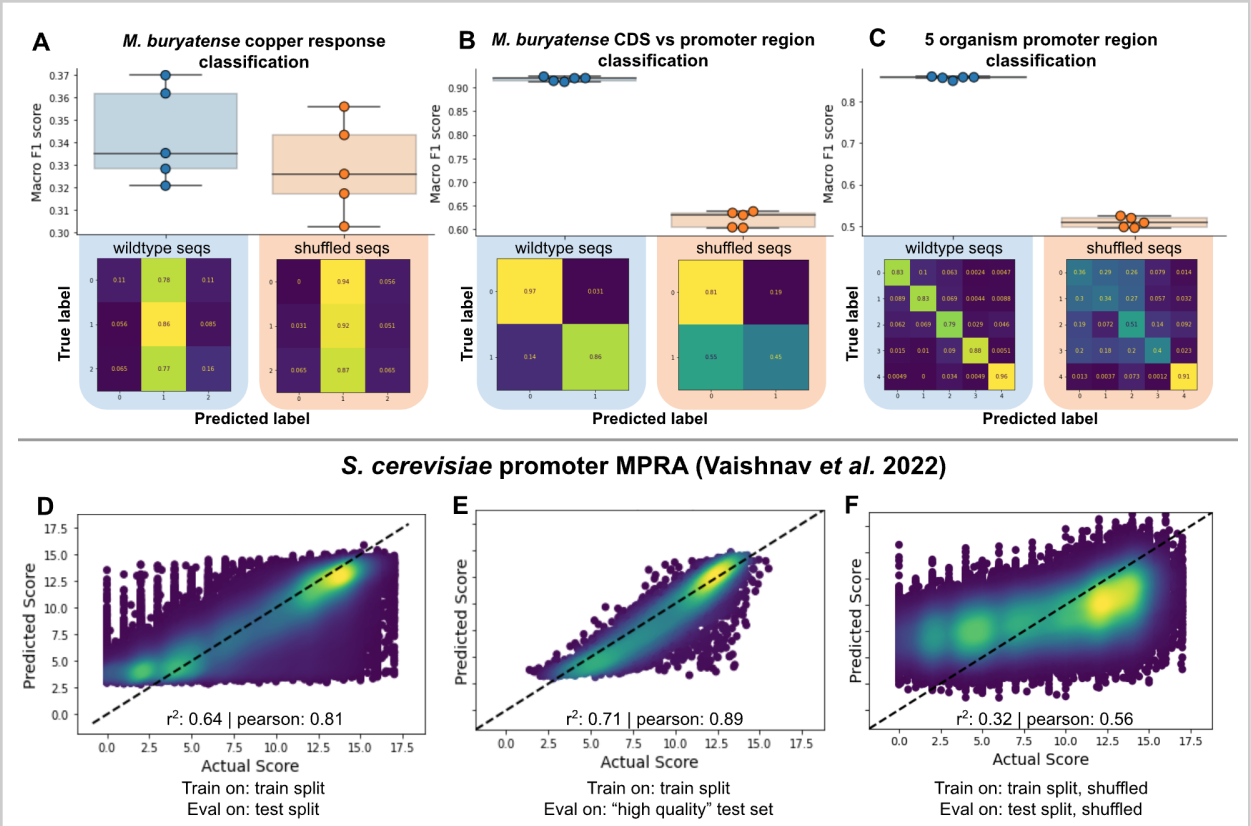


Figure 5.9. Model performance on transfer learning tasks. All models tested used the same architecture described in section 5.5, with the exception of the final prediction layer, which varied by task. **A-C)** Classification performance for models trained on wildtype sequences (blue) and shuffled sequences (orange), and evaluated on 5-folds of the dataset. A representative confusion matrix from the first fold is provided below. **A)** Task: *M. buryatense* copper response prediction. Class 0: down-regulated in High Copper; Class 1: no change in response to copper; Class 2: upregulated in response to copper. **B)** Task: *M. buryatense* coding sequence versus upstream sequence prediction. Class 0: coding sequence (CDS); Class 1: upstream region (promoter). **C)** Task: promoter region species origin prediction. Class 0: *M. buryatense*; Class 1: *E. coli*; Class 2: *B. subtilis*; Class 3: *T. maritima*; Class 4: *C. autoethanogenum*. **D-F)** Regression performance for a model trained on 1 millions random sequences and measured for promoter strength in *S. cerevisiae* in Vaishnav *et al*. 2022. Points are colored by the density of nearby points to mitigate overplotting. **D)** Model trained on sequences in the train split and evaluated on a held out test split. **E)** Model trained on sequences in the train split and evaluated on a "high quality" test set of sequences measured in a separate MPRA experiment with less measurement noise. **F)** Model trained on shuffled sequences in the train split and evaluated on a held out test split.

For our second transfer learning task, we created a dataset that included upstream regions from five different prokaryotes: *M. buryatense, E. coli, B. subtilis, T. maritima*, and *Clostridium autoethanogenum*. Most of these organisms had been investigated for constitutive promoter signals as a part of a tutorial developed for the computational framework described in Chapter 3 and showed a range of information variability in their predicted promoters. As these organisms span several distinct branches of the prokaryotic evolutionary tree and were readily available with genome annotations, we

proceed with this initial set in our transfer learning analysis. For the transfer task scenario, we similarly extracted 300bp windows upstream of the translation start site of each gene in each organism and trained models to classify the organism from which each promoter region originated. We hypothesized that this could help a model learn to identify genetic grammar patterns that evolved differently across these organisms while maintaining the specific context of upstream promoter regions, and later specialize on the *M. buryatense* copper response. We similarly found that the model architecture used performed quite well in this 5-class classification task, achieving 0.86 F1-score on the test set relative to the 0.51 for models trained on shuffled versions of the sequence (Figure 5.9.C). Notably in the shuffled models, classification for *C. autoethanogenum* was still quite successful, likely due to its extremely AT-rich genome (GC content = ~0.3) relative to the other species (Figure 5.9.C, representative confusion matrix). Both *E. coli* and *M. buryatense* have much more balanced GC content (close to ~0.5) and unsurprisingly, were most often confused with each other when the sequences were shuffled.

The third transfer learning scenario we tried was to train a model to predict expression strength from a yeast-based MPRA experiment by Vaishnav *et al.* with millions of training data sequences.[77] While *S. cerevisiae* and *M. buryatense* are more distant relatives (eukaryotic vs prokaryotic), some degree of gene expression grammar is preserved across the evolutionary tree. Especially since the MPRA training sequences are not native *S. cerevisiae* sequences but random sequences ordered from gene synthesis vendors, we suspected that features learned to predict promoter expression might still be transferable and potentially helpful during fine-tuning on the copper response task. We found that even when only training on 1 million MPRA sequences (rather than all 20 million), the model architecture was able to perform quite well. Figure 5.9.D shows the regression performance (pearson correlation of 0.81), depicting the predicted vs actual MPRA scores on a random test split from the training data. Figure 5.9.E shows the model predictions on a separate “high quality” test set, which Vaishnav *et al.* collected in a separate experiment from the training data. While it was a much smaller dataset (n = 5,298), it was a lower complexity library to reduce experimental measurement error and so predicting on this test set after training on noisier data would still be informative. On this higher quality test set, we saw better performance than on the held-out test split from training data (Pearson = 0.89). While this was lower than the correlation reported in their study (Pearson = 0.97), they used a much larger model and the full 20-million sequence training dataset. Since our goal is ultimately to fine tune a model to another task, we didn’t pursue further optimization to the transfer task as 0.89 is already quite good, especially relative to training the same model on randomized MPRA data, which was limited to a Pearson correlation of 0.56 (Figure 5.9.F).

After observing that the model architecture we defined could achieve successful prediction results on three very different training tasks, we concluded that there was not a fundamental problem with the architecture itself. It clearly is able to learn some degree of sequence signal that is different from random noise in shuffled versions of the same sequences, and thus it is able to pick up on some biological patterns relevant in each of the transfer prediction tasks. We therefore proceeded with the transfer learning experiments, initializing new models with the learned weights from each transfer task and warm-starting the training for the copper response prediction. We attempted transfer learning for the 3-class copper response classification task as well as a regression task simply aimed at predicting the log-TPM values in the High Copper condition. We evaluated model performance with class-balanced F1 score and Pearson correlation, respectively, using 5-fold cross validation across *M. buryatense* upstream promoter regions.

After evaluating each of the warm-started models, we unfortunately did not see any improvements in model performance. The average score across 5 folds for each transfer learning attempt remained about the same as the cold-started model for the copper response classification task (F1-score = 0.34), which again was quite close to the average performance on shuffled sequences (Figure 5.10.A). Similarly, the average Pearson correlation scores remained the same, if slightly worse than the cold-started models (Pearson = 0.26), though these models showed a mild improvement over models trained on shuffled sequences (Pearson = 0.19) (Figure 5.10.B)

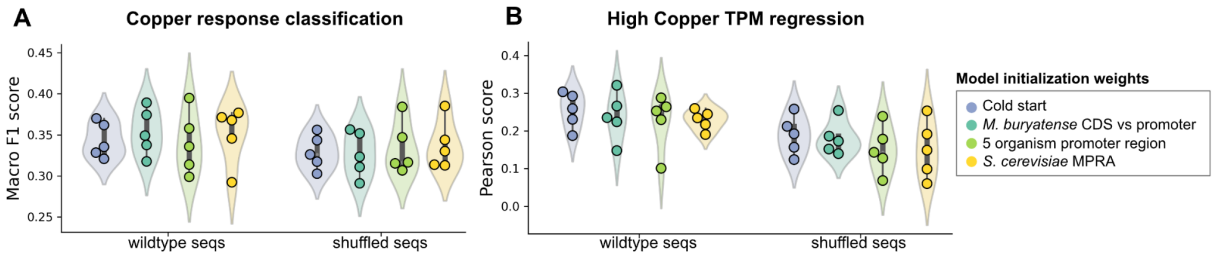


Figure 5.10. Model performance on *M. buryatense* copper response prediction after transfer learning. All models tested used the same architecture described in section 5.5, were initialized with weights from one of the successful models pre-trained on one of the tasks described in Figure 5.9, and evaluated on a held out test set over 5-folds of the *M. buryatense* promoter dataset. **A)** Performance results for the copper response classification task, measured by class-balanced F1 score for predicting Class 0: down regulated in High Copper, Class 1: no change in High Copper, Class 2: up-regulated in High Copper. **B)** Performance results for the regression task predicting $\log_2$ TPM values in the High Copper condition, measured by Pearson correlation between predicted versus measured values.

5.6. Recommendations for future deep learning analyses in non-model organisms

Unfortunately none of the transfer learning initiatives improved the learning outcomes for predicting *M. buryatense* copper response from its promoter regions. However, it was interesting to observe that the same model architecture was sufficient for achieving high performance on each of the three transfer learning tasks themselves. It is possible that none of these transfer tasks were sufficiently close to the desired task and as a result, the features learned were not relevant or tunable for the desired task. There are many other potential transfer tasks to consider and future work could certainly extend this pilot exploration.

Beyond alternative transfer learning tasks, another route to consider would be self-supervised pre-training using techniques developed for masked language modeling. In this scenario, models could be initially trained to predict "blanked out" tokens from *M. buryatense* genomic sequences, which has been shown to help models develop a more general understanding of language structure or in this case, "genomic sequence structure." Transformers in particular have revolutionized approaches to many natural language processing tasks and are actively being explored in genomic contexts as well.[134,174–176] Using the whole unlabeled genome in a self-supervised manner could unlock genetic grammar relationships that are difficult to learn from transcriptome expression measurements alone.

It is worth noting that transformers and masked language models are very computationally expensive and would thus scale the complexity and resource requirements quite dramatically. Furthermore, using computational resources with an outsized impact on greenhouse gas emissions could be detrimental to our original goal towards sustainable resource usage. Evaluating the carbon costs of any proposed transformer models with tools such as codecarbon.io would be informative for determining if incurring such an environmental cost would outweigh the benefits of better understanding the genetic grammar of any given microorganism. However, as deep learning continues to evolve at extremely rapid rates, so too will its efficiency and its accessibility, and thus it may soon become relatively easy to apply such methods in more data-limited regimes and at lower computational energy requirements.

Finally, it is possible that there is simply not enough signal present in the *M. buryatense* gene expression data for even transfer learning to overcome. Perhaps it is too noisy? Maybe some of the biological assumptions turn out to be false? For example, what if the regulatory mechanism for this particular transcriptional response is not due to a DNA binding-related signal cascade? In these cases, extending efforts with pre-training and

transfer learning will likely continue to be fruitless. If in the future resources become available to invest in adapting MPRA protocols to function in *M. buryatense*, I think that a massively high throughput exploration into this organism's expression grammar could be an exciting avenue to pursue. However, this approach would give up the goal of generalizability to other non-model organisms as such an experimental investment would have to be independently optimized for each additional organism of interest.

Overall, we suggest future researchers interested in using deep learning to model gene expression outcomes in non-model organisms attempt to estimate the information richness contained in their dataset and start their analysis with simplified synthetic tasks – such as a pair of activating and repressing motifs. If these "much easier" tasks are not achievable with the model architecture they are building at the same data scale as the dataset they have collected, then further development of their training data is required before the advantages of deep learning models can likely be useful. We hope future researchers who are exploring deep learning avenues for genomic prediction tasks consider these insights regarding data limitation and information richness, and are able to generate new information more efficiently and effectively.

Chapter 6. Summary of contributions and conclusions

6.1. Computational tools to accelerate genetic development of non-model microbial hosts

The work in this dissertation explores several computational approaches for understanding the genetic grammar in non-model microbes using RNA-seq data. We focused on the methanotroph *M. buryatense* as an example organism and used a diverse dataset that captured its transcriptional responses under various growth conditions (Chapter 2). While the dataset is unique in the specific combination of conditions it covers, RNA-seq is a relatively accessible experimental measurement and these types of varied datasets are likely common among labs aiming to characterize expression in non-standard microbes. By leveraging such collections of data to generate novel insights by analyzing gene expression patterns from multiple angles simultaneously, computational tools can accelerate tool genetic development for promising microbes with potential to serve as metabolic engineering hosts.

The first computational approach aimed to discover constitutive strong promoters by identifying genes that were in the top expressed set across all experimental conditions (Chapter 3). We derived a constitutive, strong promoter consensus for *M. buryatense*, computationally validated that it was strongly associated with promoter regions, tested a set of predicted promoters in a XylE reporter assay, and validated the sequence source of the promoter signal by scrambling various parts of the predicted sequences and re-testing their activities. This effort resulted in a suite of short (27-30bp) promoter tools that can be used to drive expression in *M. buryatense* and an open-source framework that can be readily applied to other organisms to identify strong promoters.

Our next approach moved from analyzing strong, constitutive promoters to investigating regulated expression patterns in *M. buryatense*. Specifically, we used an unsupervised machine learning method (ICA) to identify groups of genes that were independently modulated (iModulons) across the range of growth conditions in the RNA-seq data (Chapter 4). Several discovered iModulons recapitulate known regulons in *M. buryatense*, such as copper and lanthanum repressible groups, but this analysis also opened new hypotheses about putative functions for other uncharacterized response groups. This work added useful structure to the relatively under-characterized gene regulatory network of this organism and inspired future experiments to improve our understanding of its genetic grammar.

Our last approach aimed to use deep learning methods to decipher the regulatory elements underlying the highly regulated transcriptional responses observed in *M. buryatense* (Chapter 5). We were enthusiastic about the potential for these methods to automatically learn sequence features with strong influence on expression outcomes, which could be further developed into inducible promoter tools. While convolutional and recurrent neural networks had been quite successful in other genomic contexts for model organisms with large training datasets, it appears that such models are less well-suited to predicting gene expression outcomes from the relatively small genome of *M. buryatense*. After many efforts to troubleshoot class imbalance and potentially suboptimal model architectures, we conducted a meta-analysis of model performance across various data-limited regimes. MPRA datasets reduced to a similar size as the *M. buryatense* dataset performed no better than models trained on shuffled sequences, while synthetic datasets of similar size that simulated motif binding events lacked the information richness that would be required to identify even relatively simple genetic signals; both results suggested that the size of the *M. buryatense* dataset was likely a major limiting factor in model performance. We thus recommend that additional information - either in the form of more training examples or via pre-training strategies to transfer knowledge from other related tasks - is needed before the value of deep learning models can be realized for predicting expression from promoter regions in relatively small microbes.

Overall, these efforts resulted in a suite of gene expression tools, computational workflows, and insights into the applicability of deep learning methods for analyzing non-model microbial datasets more generally. The genetic tools and iModulon structure for *M. buryatense* will aid in its development as a metabolic engineering host for mitigating methane emissions. This is exciting in light of recent analysis showing that *M. buryatense* has an especially strong ability to grow at low methane concentrations typically observed in air surrounding major methane emissions sites and is a prime candidate for deploying in future methane direct air capture efforts. The characterization of deep learning models in data-limited genomic contexts can serve as a reference for others aiming to incorporate such tools into their analyses. Ideally such insights can guide future research towards pursuing more fruitful routes with their time and resources.

All together, the work throughout this dissertation supports the expanding role of computation in metabolic engineering efforts, aiming to improve the efficiency of microbial biomolecule production platforms and ultimately our ability to source materials more sustainably.

6.2. Communicating science with technical and non-technical audiences

The research described in this dissertation was communicated in a variety of outlets for a range of technical audiences. Science communication is relevant to all fields and can be effective at many levels of specificity and so I will take a final moment to describe efforts to make this work more broadly accessible.

I suspect the bulk of science communication occurs between experts in the same field, often in the same lab, as they collaborate to plan analyses and experiments on a day to day basis. Most of the same technical language is shared and thus communication of ideas can be relatively quick and smooth. Translating key insights, challenges, or ideas to experts of adjacent fields can be more cumbersome - crucial terminology is not always shared between scientific “languages,” and standard approaches in one domain are not always transferable to other contexts. Deciphering unfamiliar jargon and methods takes extra time and can be a significant drain on energy, leading to abbreviated understanding of complex topics. Especially for interdisciplinary work that requires coordination between many types of experts, ineffective communication can reduce our ability to share knowledge and make novel connections, which can hinder scientific progress more generally.

Even more challenging is communicating important findings to members of the general public, where individuals can come from such disparate technical and lived experiences that there may be little shared framework from which to draw parallels and understanding. What is not understood is difficult to trust, and without trust in a solution or course of action, how can a broader community, scientific or other, be expected to support or pursue it? Often the phrase “dumbing it down” is used to describe efforts to make scientific ideas more accessible to people with differing backgrounds; however, I think such rhetoric is a disservice, unnecessarily implying differences in intellectual ability that further isolates individuals rather than bridging understanding between them.

I do not have expertise in social psychology or policy development, but I strongly believe in the effectiveness of empathy in science communication. We should never “dumb things down,” but rather strive to clarify our work, our ideas, at varying degrees of specificity such that a given audience is readily able to engage. While the technical familiarity of diverse listeners is not always known or easy to assume, empathy and openness can go a long way to meet audiences where they are and more effectively bring them into the conversation.

Accordingly, I have strived to make the work throughout this dissertation to be intelligible to audiences with varying levels of technical backgrounds. The computational

framework in Chapter 3 was published in *ACS Synthetic Biology*[110] and the deep learning methods explored in Chapter 5 were proposed in the ICML workshop for *Tackling Climate Change with Machine Learning*.[177] However, beyond academic journals and venues, I have produced a diverse set of scientific communications for other audiences, primarily taking the form of interactive data visualizations, technical tutorials, and more general educational materials.

6.2.1. Deploying interactive data visualizations

The ability of the human visual perception system to identify coherent patterns in images rather than tables or matrices makes data visualization an invaluable asset in science communication. From summarizing broad trends to highlighting distinct anomalies, well-designed visualizations can be highly effective for transmitting insights from data. The additional ability to dynamically interrogate and filter the data is immensely helpful during initial data exploration phases of research as well as for solidifying understanding by displaying immediate responses to user-directed interactions and queries. Even an interaction as simple as a tooltip identifying a datapoint (i.e., a gene, experimental setting, motif sequence) in a scatterplot can enhance one's ability to infer relationships and often prompts additional visual designs or analyses. Furthermore, shareable interactive visualizations in the hands of other researchers with differing expertises can lead to the discovery of patterns that were not obvious or meaningful to the designer in isolation.

Each chapter in this dissertation heavily featured interactive data visualizations throughout its development, but I will highlight a few that were especially impactful for my workflows. In Chapter 2, interactive versions of PCA plots enabled inspection of specific genes or specific experimental samples across various principal components. Investigating the relative arrangement of points helped identify biological trends present, such as a dimension primarily driven by growth stress, or a dimension that primarily split experiments by growth in vials versus growth in bioreactors. Gaining this intimate familiarity with major drivers of the data was facilitated by simple mouse-over hovering, clickable legends to highlight subsets of points, and inter-plot brush filtering interactions.

In Chapter 3, one of the most useful interactive visualizations developed was a responsive parallel coordinate plot depicting expression profiles for related groups of genes across all 17 experimental conditions. Mouseover interactions emphasized nearby genes, a clickable legend enabled persistent selection of genes of interest, and a brush filter over regions of genome coordinates could additionally highlight expression profiles of co-located genes. An undergraduate researcher I mentored extended this initial concept into a full web application, incorporating gene clusters based on

transcriptomic data and GO term enrichment analysis. This tool will enable further interactive exploration of *M. buryatense* gene groups through cluster-based parallel coordinate plots and dynamically filterable tables. Other visualizations from Chapter 3 focused on exploring a tradeoff in top gene set size versus minimum TPM expression as well as upstream distances of known operons in *M. buryatense*, both of which resulted in key parameter setting decisions in our promoter identification workflow.

In Chapter 4, iModulon characterization was heavily facilitated by interactive dashboards. While we developed a few interactive charts during initial iModulon curation, the published iModulonDB implementation in javascript and highcharts provided a very sophisticated interface for exploring iModulon members, weights, and activities. The various iModulon pages were easily navigable through clicking data elements of scatter plots, barcharts, and tables, which would transport a user to a more detailed view of the clicked element. We further adapted this interface to add information such as GO terms and gene products, and organized conditions more intuitively by growth rate. Hosting these dashboards on a local server additionally provided access for other lab members to explore and pursue custom lines of inquiry that ultimately resulted in experimental proposals.

In Chapter 5, interactive data visualizations were valuable for many aspects of deep learning model troubleshooting, though these charts were not heavily featured in this final write up. Interactive parity plots allowed us to inspect which training examples the models were predicting well versus predicting poorly, prompting an analysis into sequence similarity within our promoter dataset alongside other overfitting mitigation strategies. Visualizations that displayed the results of hyper parameter searches for optimal model architectures helped us see the overriding similarity in performance across diverse model structures, suggesting that architecture was not likely our main limitation and thus continuing to vary parameters would be a less useful route to pursue. When evaluating multi-class and multi-label classification performances, interactive dashboards showing precision, recall, and F1-scores across various train/test splits and across dozens of prediction tasks prompted specific follow up analyses into class labels with seemingly better performance.

While the work in Chapter 3 was published alongside a project page with an interactive visualization gallery (https://erinhwilson.github.io/promoter-id-from-rnaseq), a selection of interactive visualizations created throughout this dissertation have been compiled and are accessible at https://erinhwilson.github.io/interactive-thesis.

6.2.2. Technical tutorials for learners in adjacent fields

While most of the interactive visualizations I created were intended for myself or other scientists in my field to gain a deeper understanding of the data presented, I additionally developed several tutorials to enhance the understanding and usability of this work.

The promoter identification framework described in Chapter 3 was published with a suite of interactive tutorials demonstrating key analyses throughout the computational workflow. These include explorations of the expression data and intra-operon gene distances, inspection of Bioprospector output files that were parsed to form a consensus motif, computational validation of the consensus motif, and a demonstration of analyzing the framework outputs for other organisms and comparing the results. Similarly, when investigating model performance for gene expression prediction from promoter sequences across data-limited regimes in Chapter 5, I compiled streamlined notebooks outlining analysis steps for both the MPRA reduction experiment and the synthetic sequence motif prediction task. These materials contain documentation and explanations in readable prose, and should help researchers interested in using these tools to more effectively apply them to organisms or datasets in their own domain.

While learning to develop deep learning models in Chapter 5, I noticed a gap in online learning materials. There existed both basic PyTorch tutorials for using CNNs to process image data and also sophisticated deep learning repositories for making predictions from genomic sequence data. However, there were no tutorials that were both targeted for beginners and DNA sequences. To fill this need, I created an example from our synthetic sequence prediction analysis and distilled the key concepts into a beginner-friendly tutorial – “Modeling DNA sequences with PyTorch” – showing how to adapt the PyTorch deep learning framework for DNA sequence inputs (Figure 6.1.A). Its publication in *Towards Data Science* has enabled the tutorial to reach over six-thousand viewers beyond academic journals, including machine learning practitioners growing their knowledge of biology, biologists just starting out with deep learning tools, and general data science enthusiasts on the internet (Figure 6.1.C).

6.2.3. Educational materials for general audiences

Because day-to-day research activities are often so zoomed into deciphering complex details at the frontier of our collective knowledge, why these details are worth investigating, supporting, or even funding is sometimes difficult to convey to external groups. Impactful scientific solutions are more likely to garner support from broader communities if they are relatable and understandable, thus effective scientific communication with the general public is extremely important.

I have remained incredibly excited about the possibilities of synthetic biology to make a difference in global sustainability, however discussing the field of genetic engineering comes with significant baggage. For many valid reasons, many people do not trust genetic engineering, whether it be negative consequences with some bad actors in the space or fear of unpredictable consequences of new technology. While “evil scientist” tropes common in the media do not help with negative associations of genetically modified organisms, I was motivated to communicate ways in which genetic engineering can actually be quite helpful, especially with respect to sustainability. I’ve found that many people have not heard of metabolic engineering for sustainable biomolecule production, and so writing “The Light Side of Genetic Engineering” for the publication *OneZero* was a highly enjoyable effort to share an alternative view to the commonly conveyed GMO narrative (Figure 6.1.B). While this is but one article in a sea of media content, I wish to continue using accessible prose to connect with audiences beyond the scientists and engineers I most frequently encounter in my everyday work. Ultimately, if our scientific innovations cannot translate out of the lab and into the public domain, our ability to meaningfully impact climate change will be negligible.

Finally, while engaging with general audiences about scientific solutions can certainly aid their adoption, making scientific ideas accessible to young people can cultivate an inspired energy that takes root more broadly within the next generation. I’ve taken several opportunities to introduce concepts from metabolic and genetic engineering to elementary and middle school learners, creating interactive, tactile activities with magnets and cartoon organisms with uniquely shaped “DNA” puzzle pieces (Figure 6.1.D). While the activity mechanics are dramatically simplified relative to the genetic manipulations that go into actual microbe engineering work, they are clarified at an appropriate level such that younger audiences remain engaged with the core ideas and, hopefully, are inspired to learn more about the possibilities of such ideas to protect the environment.

A quote from Mrs. Terwilliger, the founder of a wildlife rescue hospital and educational program in San Rafael, CA, is particularly meaningful for me: “Teach children to love Nature. They’ll take care of what they love.” Striving to grow societal support through inspiration, especially in future generations, is perhaps an underlooked avenue in scientific endeavors, but one I plan to continue pursuing throughout my career.

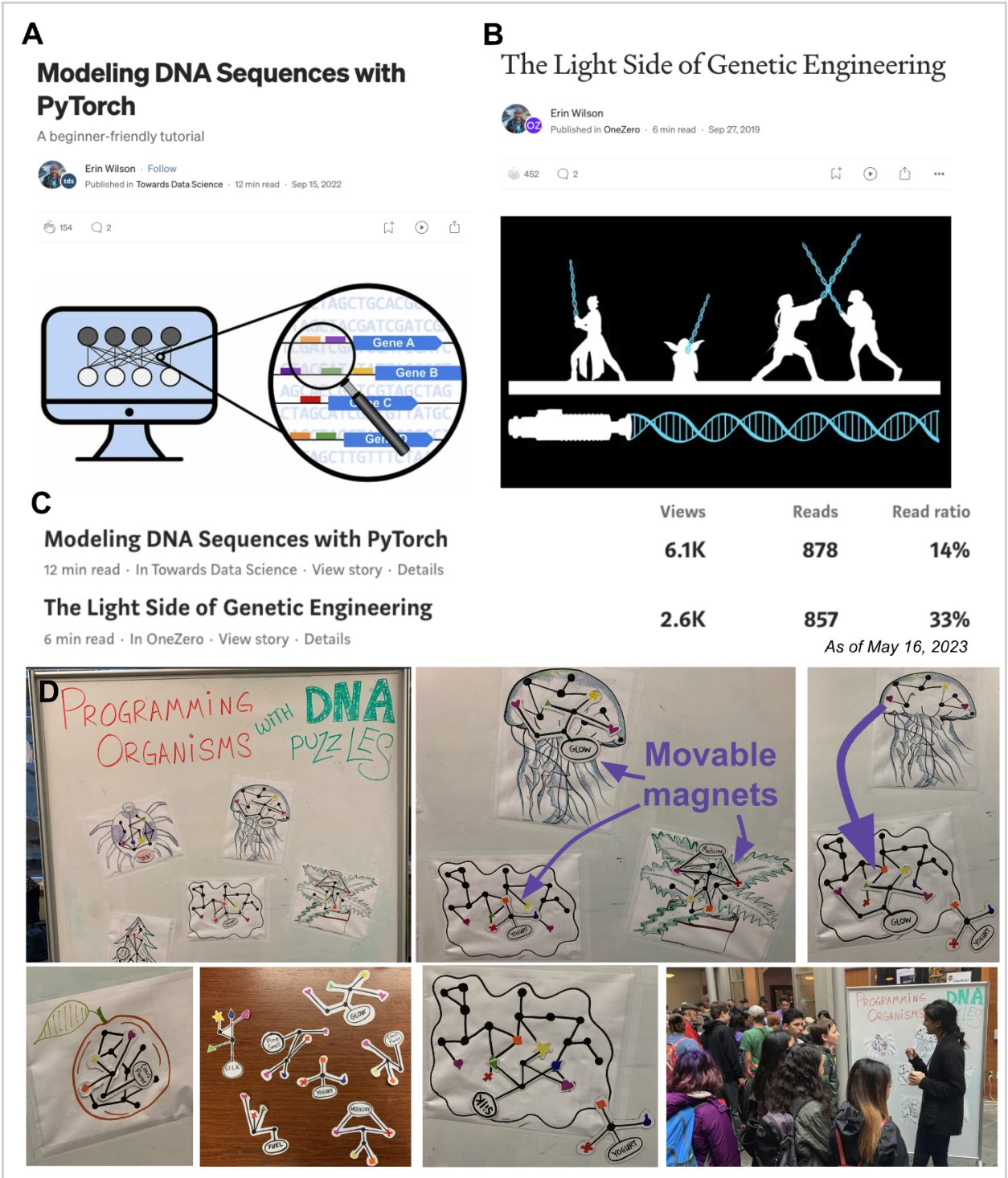


Figure 6.1. Science communications for varying technical audiences. **A)** "Modeling DNA Sequences with PyTorch." A beginner-friendly tutorial published in *Towards Data Science*, a Medium publication (https://towardsdatascience.com/modeling-dna-sequences-with-pytorch-de28b0a05036). **B)** "The Light Side of Genetic Engineering." An introduction to metabolic engineering published in *OneZero*, a Medium publication (https://onezero.medium.com/the-light-side-of-genetic-engineering-a65c5863b4d8). **C)** Audience view and read statistics as of May 16, 2023. **D)** "Programming organisms with DNA puzzles." An interactive activity for elementary and middle schoolers. The activity features a suite of

organisms that have the DNA to make a variety of useful molecules (spider silk, biofuel, citrus flavor, pine smell, medicine, glow proteins, yogurt), represented as movable magnets that fit into the puzzle structure inside its host organism (network of colored shapes). As participants try moving around the puzzle pieces (as a representation of transferring DNA instructions between organisms), they notice that the “bacterium” puzzle pocket is flexible enough for all the other “products” to fit, indicating its relative ease of engineering.

6.3. Conclusion

Upon explaining the major goals of my research, many listeners are surprised to hear it was done in a computer science department. Studying bacterial DNA sequences and manipulating organism genomes certainly *sounds* more like biology. Noting that the vast amounts of biological data being collected requires the aid of computational methods to process and analyze it often helps clarify my membership in the school of computer science and engineering; but truthfully, this work “could fit” within the research goals of any number of departments. Genome sciences, chemical engineering, bioengineering, microbiology, and molecular engineering are perhaps the most obvious candidates, but others are quite relevant, too.

More accurately, this work is supported and fueled by innovations made in many different fields. The convergence of disparate ideas and methods – from deep learning theory proposed decades ago, next generation sequencing technology, advances in GPU hardware, optimized methanotroph laboratory cultivation, synthetic biology tools for efficient microbial genome editing, and too many scientific measurement technologies to name – all enabled our computational explorations into the *M. buryatense* genetic grammar. I hope the insights generated throughout this dissertation contribute to continued interdisciplinary growth within and between scientific fields as we work to shift human societies towards a stable relationship with our environment.

References

(1) Cameron, D. E.; Bashor, C. J.; Collins, J. J. A Brief History of Synthetic Biology. *Nat. Rev. Microbiol.* **2014**, *12* (5), 381–390. https://doi.org/10.1038/nrmicro3239.

(2) de Lorenzo, V.; Prather, K. L.; Chen, G.; O'Day, E.; von Kameke, C.; Oyarzún, D. A.; Hosta‐Rigau, L.; Alsafar, H.; Cao, C.; Ji, W.; Okano, H.; Roberts, R. J.; Ronaghi, M.; Yeung, K.; Zhang, F.; Lee, S. Y. The Power of Synthetic Biology for Bioproduction, Remediation and Pollution Control. *EMBO Rep.* **2018**, *19* (4), e45658. https://doi.org/10.15252/embr.201745658.

(3) Cravens, A.; Payne, J.; Smolke, C. D. Synthetic Biology Strategies for Microbial Biosynthesis of Plant Natural Products. *Nat. Commun.* **2019**, *10* (1), 2142. https://doi.org/10.1038/s41467-019-09848-w.

(4) Pickens, L. B.; Tang, Y.; Chooi, Y.-H. Metabolic Engineering for the Production of Natural Products. *Annu. Rev. Chem. Biomol. Eng.* **2011**, *2*, 211–236. https://doi.org/10.1146/annurev-chembioeng-061010-114209.

(5) Nielsen, J.; Keasling, J. D. Engineering Cellular Metabolism. *Cell* **2016**, *164* (6), 1185–1197. https://doi.org/10.1016/j.cell.2016.02.004.

(6) Woolston, B. M.; Edgar, S.; Stephanopoulos, G. Metabolic Engineering: Past and Future. *Annu. Rev. Chem. Biomol. Eng.* **2013**, *4* (1), 259–288. https://doi.org/10.1146/annurev-chembioeng-061312-103312.

(7) Voigt, C. A. Synthetic Biology 2020–2030: Six Commercially-Available Products That Are Changing Our World. *Nat. Commun.* **2020**, *11* (1), 6379. https://doi.org/10.1038/s41467-020-20122-2.

(8) Department of Defense Biomanufacturing Strategy.

(9) Scown, C. D.; Keasling, J. D. Sustainable Manufacturing with Synthetic Biology. *Nat. Biotechnol.* **2022**, *40* (3), 304–307. https://doi.org/10.1038/s41587-022-01248-8.

(10) Macfarlane, N. B. W.; Adams, J.; Bennett, E. L.; Brooks, T. M.; Delborne, J. A.; Eggermont, H.; Endy, D.; Esvelt, K. M.; Kolodziejczyk, B.; Kuiken, T.; Oliva, M. J.; Peña Moreno, S.; Slobodian, L.; Smith, R. B.; Thizy, D.; Tompkins, D. M.; Wei, W.; Redford, K. H. Direct and Indirect Impacts of Synthetic Biology on Biodiversity Conservation. *iScience* **2022**, *25* (11), 105423. https://doi.org/10.1016/j.isci.2022.105423.

(11) Benjamin, K. R.; Silva, I. R.; Cherubim, J. P.; McPhee, D.; Paddon, C. J.; Benjamin, K. R.; Silva, I. R.; Cherubim, J. P.; McPhee, D.; Paddon, C. J. Developing Commercial Production of Semi-Synthetic Artemisinin, and of β-Farnesene, an Isoprenoid Produced by Fermentation of Brazilian Sugar. *J. Braz. Chem. Soc.* **2016**, *27* (8), 1339–1345. https://doi.org/10.5935/0103-5053.20160119.

(12) Meadows, A. L.; Hawkins, K. M.; Tsegaye, Y.; Antipov, E.; Kim, Y.; Raetz, L.; Dahl, R. H.; Tai, A.; Mahatdejkul-Meadows, T.; Xu, L.; Zhao, L.; Dasika, M. S.; Murarka, A.; Lenihan, J.; Eng, D.; Leng, J. S.; Liu, C.-L.; Wenger, J. W.; Jiang, H.; Chao, L.; Westfall, P.; Lai, J.; Ganesan, S.; Jackson, P.; Mans, R.; Platt, D.; Reeves, C. D.; Saija, P. R.; Wichmann, G.; Holmes, V. F.; Benjamin, K.; Hill, P. W.; Gardner, T. S.; Tsong, A. E. Rewriting Yeast Central Carbon Metabolism for Industrial Isoprenoid Production. *Nature* **2016**, *537* (7622), 694–697. https://doi.org/10.1038/nature19769.

(13) Ro, D.-K.; Paradise, E. M.; Ouellet, M.; Fisher, K. J.; Newman, K. L.; Ndungu, J. M.; Ho, K. A.; Eachus, R. A.; Ham, T. S.; Kirby, J.; Chang, M. C. Y.; Withers, S. T.; Shiba, Y.; Sarpong, R.; Keasling, J. D. Production of the Antimalarial Drug Precursor Artemisinic Acid in Engineered Yeast. *Nature* **2006**, *440* (7086), 940–943. https://doi.org/10.1038/nature04640.

(14) Paddon, C. J.; Westfall, P. J.; Pitera, D. J.; Benjamin, K.; Fisher, K.; McPhee, D.; Leavell,

M. D.; Tai, A.; Main, A.; Eng, D.; Polichuk, D. R.; Teoh, K. H.; Reed, D. W.; Treynor, T.; Lenihan, J.; Jiang, H.; Fleck, M.; Bajad, S.; Dang, G.; Dengrove, D.; Diola, D.; Dorin, G.; Ellens, K. W.; Fickes, S.; Galazzo, J.; Gaucher, S. P.; Geistlinger, T.; Henry, R.; Hepp, M.; Horning, T.; Iqbal, T.; Kizer, L.; Lieu, B.; Melis, D.; Moss, N.; Regentin, R.; Secrest, S.; Tsuruta, H.; Vazquez, R.; Westblade, L. F.; Xu, L.; Yu, M.; Zhang, Y.; Zhao, L.; Lievense, J.; Covello, P. S.; Keasling, J. D.; Reiling, K. K.; Renninger, N. S.; Newman, J. D. High-Level Semi-Synthetic Production of the Potent Antimalarial Artemisinin. *Nature* **2013**, *496* (7446), 528–532. https://doi.org/10.1038/nature12051.

(15) Yim, H.; Haselbeck, R.; Niu, W.; Pujol-Baxley, C.; Burgard, A.; Boldt, J.; Khandurina, J.; Trawick, J. D.; Osterhout, R. E.; Stephen, R.; Estadilla, J.; Teisan, S.; Schreyer, H. B.; Andrae, S.; Yang, T. H.; Lee, S. Y.; Burk, M. J.; Van Dien, S. Metabolic Engineering of Escherichia Coli for Direct Production of 1,4-Butanediol. *Nat. Chem. Biol.* **2011**, *7* (7), 445–452. https://doi.org/10.1038/nchembio.580.

(16) Nakamura, C. E.; Whited, G. M. Metabolic Engineering for the Microbial Production of 1,3-Propanediol. *Curr. Opin. Biotechnol.* **2003**, *14* (5), 454–459. https://doi.org/10.1016/j.copbio.2003.08.005.

(17) Widmaier, D. M.; Tullman‐Ercek, D.; Mirsky, E. A.; Hill, R.; Govindarajan, S.; Minshull, J.; Voigt, C. A. Engineering the Salmonella Type III Secretion System to Export Spider Silk Monomers. *Mol. Syst. Biol.* **2009**, *5* (1), 309. https://doi.org/10.1038/msb.2009.62.

(18) Shankar, S.; Hoyt, M. A. Expression Constructs and Methods of Genetically Engineering Methylotrophic Yeast. US20170349906A1, December 7, 2017. https://patents.google.com/patent/US20170349906A1/en (accessed 2021-11-27).

(19) Elena, C.; Ravasi, P.; Castelli, M.; Peiru, S.; Menzella, H. Expression of Codon Optimized Genes in Microbial Systems: Current Industrial Applications and Perspectives. *Front. Microbiol.* **2014**, *5*.

(20) Bervoets, I.; Charlier, D. Diversity, Versatility and Complexity of Bacterial Gene Regulation Mechanisms: Opportunities and Drawbacks for Applications in Synthetic Biology. *FEMS Microbiol. Rev.* **2019**. https://doi.org/10.1093/femsre/fuz001.

(21) Ren, G.-X.; Guo, X.-P.; Sun, Y.-C. Regulatory 3′ Untranslated Regions of Bacterial MRNAs. *Front. Microbiol.* **2017**, *8*. https://doi.org/10.3389/fmicb.2017.01276.

(22) Paget, M. S. B.; Helmann, J. D. The Sigma70 Family of Sigma Factors. *Genome Biol.* **2003**, *4* (1), 203. https://doi.org/10.1186/gb-2003-4-1-203.

(23) Feklístov, A.; Sharon, B. D.; Darst, S. A.; Gross, C. A. Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annu. Rev. Microbiol.* **2014**, *68*, 357–376. https://doi.org/10.1146/annurev-micro-092412-155737.

(24) Browning, D. F.; Butala, M.; Busby, S. J. W. Bacterial Transcription Factors: Regulation by Pick “N” Mix. *J. Mol. Biol.* **2019**, *431* (20), 4067–4077. https://doi.org/10.1016/j.jmb.2019.04.011.

(25) Claassens, N. J.; Sousa, D. Z.; dos Santos, V. A. P. M.; de Vos, W. M.; van der Oost, J. Harnessing the Power of Microbial Autotrophy. *Nat. Rev. Microbiol.* **2016**, *14* (11), 692–706. https://doi.org/10.1038/nrmicro.2016.130.

(26) Blombach, B.; Grünberger, A.; Centler, F.; Wierckx, N.; Schmid, J. Exploiting Unconventional Prokaryotic Hosts for Industrial Biotechnology. *Trends Biotechnol.* **2021**, S0167779921001906. https://doi.org/10.1016/j.tibtech.2021.08.003.

(27) Puri, A. W.; Owen, S.; Chu, F.; Chavkin, T.; Beck, D. A. C.; Kalyuzhnaya, M. G.; Lidstrom, M. E. Genetic Tools for the Industrially Promising Methanotroph Methylomicrobium Buryatense. *Appl Env. Microbiol* **2015**, *81* (5), 1775–1781. https://doi.org/10.1128/AEM.03795-14.

(28) Portela, R. M. C.; Vogl, T.; Kniely, C.; Fischer, J. E.; Oliveira, R.; Glieder, A. Synthetic Core Promoters as Universal Parts for Fine-Tuning Expression in Different Yeast Species. *ACS Synth. Biol.* **2017**, *6* (3), 471–484. https://doi.org/10.1021/acssynbio.6b00178.

(29) Browning, D. F.; Busby, S. J. W. The Regulation of Bacterial Transcription Initiation. *Nat. Rev. Microbiol.* **2004**, *2* (1), 57–65. https://doi.org/10.1038/nrmicro787.
(30) Kosuri, S.; Goodman, D. B.; Cambray, G.; Mutalik, V. K.; Gao, Y.; Arkin, A. P.; Endy, D.; Church, G. M. Composability of Regulatory Sequences Controlling Transcription and Translation in Escherichia Coli. *Proc. Natl. Acad. Sci.* **2013**, *110* (34), 14024–14029. https://doi.org/10.1073/pnas.1301301110.
(31) Saunois, M.; Stavert, A. R.; Poulter, B.; Bousquet, P.; Canadell, J. G.; Jackson, R. B.; Raymond, P. A.; Dlugokencky, E. J.; Houweling, S.; Patra, P. K.; Ciais, P.; Arora, V. K.; Bastviken, D.; Bergamaschi, P.; Blake, D. R.; Brailsford, G.; Bruhwiler, L.; Carlson, K. M.; Carrol, M.; Castaldi, S.; Chandra, N.; Crevoisier, C.; Crill, P. M.; Covey, K.; Curry, C. L.; Etiope, G.; Frankenberg, C.; Gedney, N.; Hegglin, M. I.; Höglund-Isaksson, L.; Hugelius, G.; Ishizawa, M.; Ito, A.; Janssens-Maenhout, G.; Jensen, K. M.; Joos, F.; Kleinen, T.; Krummel, P. B.; Langenfelds, R. L.; Laruelle, G. G.; Liu, L.; Machida, T.; Maksyutov, S.; McDonald, K. C.; McNorton, J.; Miller, P. A.; Melton, J. R.; Morino, I.; Müller, J.; Murguia-Flores, F.; Naik, V.; Niwa, Y.; Noce, S.; O'Doherty, S.; Parker, R. J.; Peng, C.; Peng, S.; Peters, G. P.; Prigent, C.; Prinn, R.; Ramonet, M.; Regnier, P.; Riley, W. J.; Rosentreter, J. A.; Segers, A.; Simpson, I. J.; Shi, H.; Smith, S. J.; Steele, L. P.; Thornton, B. F.; Tian, H.; Tohjima, Y.; Tubiello, F. N.; Tsuruta, A.; Viovy, N.; Voulgarakis, A.; Weber, T. S.; van Weele, M.; van der Werf, G. R.; Weiss, R. F.; Worthy, D.; Wunch, D.; Yin, Y.; Yoshida, Y.; Zhang, W.; Zhang, Z.; Zhao, Y.; Zheng, B.; Zhu, Q.; Zhu, Q.; Zhuang, Q. The Global Methane Budget 2000–2017. *Earth Syst. Sci. Data* **2020**, *12* (3), 1561–1623. https://doi.org/10.5194/essd-12-1561-2020.
(32) Hanson, R. S.; Hanson, T. E. Methanotrophic Bacteria. *Microbiol. Rev.* **1996**, *60* (2), 439–471.
(33) Chistoserdova, L. Modularity of Methylotrophy, Revisited. *Environ. Microbiol.* **2011**, *13* (10), 2603–2622. https://doi.org/10.1111/j.1462-2920.2011.02464.x.
(34) Marcellin, E.; Behrendorff, J. B.; Nagaraju, S.; DeTissera, S.; Segovia, S.; Palfreyman, R. W.; Daniell, J.; Licona-Cassani, C.; Quek, L.; Speight, R.; Hodson, M. P.; Simpson, S. D.; Mitchell, W. P.; Köpke, M.; Nielsen, L. K. Low Carbon Fuels and Commodity Chemicals from Waste Gases – Systematic Approach to Understand Energy Metabolism in a Model Acetogen. *Green Chem.* **2016**, *18* (10), 3020–3028. https://doi.org/10.1039/C5GC02708J.
(35) Kalyuzhnaya, M. G.; Puri, A. W.; Lidstrom, M. E. Metabolic Engineering in Methanotrophic Bacteria. *Metab. Eng.* **2015**, *29*, 142–152. https://doi.org/10.1016/j.ymben.2015.03.010.
(36) François, J. M.; Lachaux, C.; Morin, N. Synthetic Biology Applied to Carbon Conservative and Carbon Dioxide Recycling Pathways. *Front. Bioeng. Biotechnol.* **2020**, *7*.
(37) Jiang, W.; Hernández Villamor, D.; Peng, H.; Chen, J.; Liu, L.; Haritos, V.; Ledesma-Amaro, R. Metabolic Engineering Strategies to Enable Microbial Utilization of C1 Feedstocks. *Nat. Chem. Biol.* **2021**, *17* (8), 845–855. https://doi.org/10.1038/s41589-021-00836-0.
(38) Shindell, D.; Kuylenstierna, J. C. I.; Vignati, E.; Dingenen, R. van; Amann, M.; Klimont, Z.; Anenberg, S. C.; Muller, N.; Janssens-Maenhout, G.; Raes, F.; Schwartz, J.; Faluvegi, G.; Pozzoli, L.; Kupiainen, K.; Höglund-Isaksson, L.; Emberson, L.; Streets, D.; Ramanathan, V.; Hicks, K.; Oanh, N. T. K.; Milly, G.; Williams, M.; Demkine, V.; Fowler, D. Simultaneously Mitigating Near-Term Climate Change and Improving Human Health and Food Security. *Science* **2012**, *335* (6065), 183–189. https://doi.org/10.1126/science.1210026.
(39) Myhre, G.; Shindell, D.; Bréon, F.-M.; Collins, W.; Fuglestvedt, J.; Huang, J.; Koch, D.; Lamarque, J.-F.; Lee, D.; Mendoza, B.; Nakajima, T.; Robock, A.; Stephens, G.; Zhang, H.; Aamaas, B.; Boucher, O.; Dalsøren, S. B.; Daniel, J. S.; Forster, P.; Granier, C.; Haigh, J.; Hodnebrog, Ø.; Kaplan, J. O.; Marston, G.; Nielsen, C. J.; O'Neill, B. C.; Peters, G. P.; Pongratz, J.; Ramaswamy, V.; Roth, R.; Rotstayn, L.; Smith, S. J.; Stevenson, D.; Vernier,

J.-P.; Wild, O.; Young, P.; Jacob, D.; Ravishankara, A. R.; Shine, K. 8 Anthropogenic and Natural Radiative Forcing.
(40) Pratt, C.; Tate, K. Mitigating Methane: Emerging Technologies To Combat Climate Change's Second Leading Contributor. *Environ. Sci. Technol.* **2018**, *52* (11), 6084–6097. https://doi.org/10.1021/acs.est.7b04711.
(41) Abernethy, S.; O'Connor, F. M.; Jones, C. D.; Jackson, R. B. Methane Removal and the Proportional Reductions in Surface Temperature and Ozone. *Philos. Transact. A Math. Phys. Eng. Sci. 379* (2210), 20210104. https://doi.org/10.1098/rsta.2021.0104.
(42) Gilman, A.; Fu, Y.; Hendershott, M.; Chu, F.; Puri, A. W.; Smith, A. L.; Pesesky, M.; Lieberman, R.; Beck, D. A. C.; Lidstrom, M. E. Oxygen-Limited Metabolism in the Methanotroph Methylomicrobium Buryatense 5GB1C. *PeerJ* **2017**, *5*, e3945. https://doi.org/10.7717/peerj.3945.
(43) de la Torre, A.; Metivier, A.; Chu, F.; Laurens, L. M. L.; Beck, D. A. C.; Pienkos, P. T.; Lidstrom, M. E.; Kalyuzhnaya, M. G. Genome-Scale Metabolic Reconstructions and Theoretical Investigation of Methane Conversion in Methylomicrobium Buryatense Strain 5G(B1). *Microb. Cell Factories* **2015**, *14*. https://doi.org/10.1186/s12934-015-0377-3.
(44) Fu, Y.; Li, Y.; Lidstrom, M. The Oxidative TCA Cycle Operates during Methanotrophic Growth of the Type I Methanotroph Methylomicrobium Buryatense 5GB1. *Metab. Eng.* **2017**, *42*, 43–51. https://doi.org/10.1016/j.ymben.2017.05.003.
(45) Fu, Y.; He, L.; Reeve, J.; Beck, D. A. C.; Lidstrom, M. E. Core Metabolism Shifts during Growth on Methanol versus Methane in the Methanotroph Methylomicrobium Buryatense 5GB1. *mBio* **2019**, *10* (2), e00406-19. https://doi.org/10.1128/mBio.00406-19.
(46) US Department of Commerce, N. *Global Monitoring Laboratory - Carbon Cycle Greenhouse Gases*. https://gml.noaa.gov/ccgg/trends_ch4/ (accessed 2023-04-03).
(47) Svensson, B. H. A National Landfill Methane Budget for Sweden Based on Field Measurements, and an Evaluation of IPCC Models. **2009**, *61* (2), 424. https://doi.org/10.1111/j.1600-0889.2008.00409.x.
(48) Irakulis-Loitxate, I.; Guanter, L.; Liu, Y.-N.; Varon, D. J.; Maasakkers, J. D.; Zhang, Y.; Chulakadabba, A.; Wofsy, S. C.; Thorpe, A. K.; Duren, R. M.; Frankenberg, C.; Lyon, D. R.; Hmiel, B.; Cusworth, D. H.; Zhang, Y.; Segl, K.; Gorroño, J.; Sánchez-García, E.; Sulprizio, M. P.; Cao, K.; Zhu, H.; Liang, J.; Li, X.; Aben, I.; Jacob, D. J. Satellite-Based Survey of Extreme Methane Emissions in the Permian Basin. *Sci. Adv.* **2021**, *7* (27), eabf4507. https://doi.org/10.1126/sciadv.abf4507.
(49) Wu, S.; Li, S.; Zou, Z.; Hu, T.; Hu, Z.; Liu, S.; Zou, J. High Methane Emissions Largely Attributed to Ebullitive Fluxes from a Subtropical River Draining a Rice Paddy Watershed in China. *Environ. Sci. Technol.* **2019**, *53* (7), 3499–3507. https://doi.org/10.1021/acs.est.8b05286.
(50) Yoon, S.; Carey, J. N.; Semrau, J. D. Feasibility of Atmospheric Methane Removal Using Methanotrophic Biotrickling Filters. *Appl. Microbiol. Biotechnol.* **2009**, *83* (5), 949–956. https://doi.org/10.1007/s00253-009-1977-9.
(51) Crick, F. Central Dogma of Molecular Biology. *Nature* **1970**, *227* (5258), 561–563. https://doi.org/10.1038/227561a0.
(52) van Dam, S.; Võsa, U.; van der Graaf, A.; Franke, L.; de Magalhães, J. P. Gene Co-Expression Analysis for Functional Classification and Gene–Disease Predictions. *Brief. Bioinform.* **2018**, *19* (4), 575–592. https://doi.org/10.1093/bib/bbw139.
(53) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* **2009**, *10* (1), 57–63. https://doi.org/10.1038/nrg2484.
(54) Croucher, N. J.; Thomson, N. R. Studying Bacterial Transcriptomes Using RNA-Seq. *Curr. Opin. Microbiol.* **2010**, *13* (5), 619–624. https://doi.org/10.1016/j.mib.2010.09.009.
(55) De Smet, R.; Marchal, K. Advantages and Limitations of Current Network Inference Methods. *Nat. Rev. Microbiol.* **2010**, *8* (10), 717–729. https://doi.org/10.1038/nrmicro2419.

(56) Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinforma. Oxf. Engl.* **2009**, *25* (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
(57) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinforma. Oxf. Engl.* **2009**, *25* (16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
(58) Anders, S.; Pyl, P. T.; Huber, W. HTSeq--a Python Framework to Work with High-Throughput Sequencing Data. *Bioinforma. Oxf. Engl.* **2015**, *31* (2), 166–169. https://doi.org/10.1093/bioinformatics/btu638.
(59) Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M. W.; Gaffney, D. J.; Elo, L. L.; Zhang, X.; Mortazavi, A. A Survey of Best Practices for RNA-Seq Data Analysis. *Genome Biol.* **2016**, *17*. https://doi.org/10.1186/s13059-016-0881-8.
(60) Eddy, S. R. Non–Coding RNA Genes and the Modern RNA World. *Nat. Rev. Genet.* **2001**, *2* (12), 919–929. https://doi.org/10.1038/35103511.
(61) Janssen, B. D.; Hayes, C. S. The TmRNA Ribosome Rescue System. *Adv. Protein Chem. Struct. Biol.* **2012**, *86*, 151–191. https://doi.org/10.1016/B978-0-12-386497-0.00005-0.
(62) Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* **2013**, *29* (1), 15–21. https://doi.org/10.1093/bioinformatics/bts635.
(63) Sastry, A. V.; Gao, Y.; Szubin, R.; Hefner, Y.; Xu, S.; Kim, D.; Choudhary, K. S.; Yang, L.; King, Z. A.; Palsson, B. O. The Escherichia Coli Transcriptome Mostly Consists of Independently Regulated Modules. *Nat. Commun.* **2019**, *10* (1), 5536. https://doi.org/10.1038/s41467-019-13483-w.
(64) Rychel, K.; Sastry, A. V.; Palsson, B. O. Machine Learning Uncovers Independently Regulated Modules in the Bacillus Subtilis Transcriptome. *Nat. Commun.* **2020**, *11* (1), 6338. https://doi.org/10.1038/s41467-020-20153-9.
(65) Umemura, M.; Kuriiwa, K.; Dao, L. V.; Okuda, T.; Terai, G. Promoter Tools for Further Development of Aspergillus Oryzae as a Platform for Fungal Secondary Metabolite Production. *Fungal Biol. Biotechnol.* **2020**, *7* (1), 3. https://doi.org/10.1186/s40694-020-00093-1.
(66) Li, S.; Wang, J.; Li, X.; Yin, S.; Wang, W.; Yang, K. Genome-Wide Identification and Evaluation of Constitutive Promoters in Streptomycetes. *Microb. Cell Factories* **2015**. https://doi.org/10.1186/s12934-015-0351-0.
(67) Luo, Y.; Zhang, L.; Barton, K. W.; Zhao, H. Systematic Identification of a Panel of Strong Constitutive Promoters from *Streptomyces Albus*. *ACS Synth. Biol.* **2015**, *4* (9), 1001–1010. https://doi.org/10.1021/acssynbio.5b00016.
(68) Ouyang, Q.; Wang, X.; Zhang, N.; Zhong, L.; Liu, J.; Ding, X.; Zhang, Y.; Bian, X. Promoter Screening Facilitates Heterologous Production of Complex Secondary Metabolites in Burkholderiales Strains. *ACS Synth. Biol.* **2020**, *9* (2), 457–460. https://doi.org/10.1021/acssynbio.9b00459.
(69) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao,

J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.
(70) Zhu, X.; Vondrick, C.; Fowlkes, C. C.; Ramanan, D. Do We Need More Training Data? *Int. J. Comput. Vis.* **2016**, *119* (1), 76–92. https://doi.org/10.1007/s11263-015-0812-2.
(71) Bailly, A.; Blanc, C.; Francis, É.; Guillotin, T.; Jamal, F.; Wakim, B.; Roy, P. Effects of Dataset Size and Interactions on the Prediction Performance of Logistic Regression and Deep Learning Models. *Comput. Methods Programs Biomed.* **2022**, *213*, 106504. https://doi.org/10.1016/j.cmpb.2021.106504.
(72) Cuperus, J. T.; Groves, B.; Kuchina, A.; Rosenberg, A. B.; Jojic, N.; Fields, S.; Seelig, G. Deep Learning of the Regulatory Grammar of Yeast 5' Untranslated Regions from 500,000 Random Sequences. *Genome Res.* **2017**, *27* (12), 2015–2024. https://doi.org/10.1101/gr.224964.117.
(73) Sample, P. J.; Wang, B.; Reid, D. W.; Presnyak, V.; McFadyen, I.; Morris, D. R.; Seelig, G. Human 5' UTR Design and Variant Effect Prediction from a Massively Parallel Translation Assay. *Nat. Biotechnol.* **2019**, *37* (7), 803–809. https://doi.org/10.1038/s41587-019-0164-5.
(74) Agarwal, V.; Inoue, F.; Schubach, M.; Martin, B. K.; Mohan, P.; Zhang, Z.; Sohota, A.; Noble, W. S.; Yardimci, G. G.; Kircher, M.; Shendure, J.; Ahituv, N. Massively Parallel Characterization of Transcriptional Regulatory Elements in Three Diverse Human Cell Types.
(75) Bogard, N.; Linder, J.; Rosenberg, A. B.; Seelig, G. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **2019**, *0* (0). https://doi.org/10.1016/j.cell.2019.04.046.
(76) de Boer, C. G.; Vaishnav, E. D.; Sadeh, R.; Abeyta, E. L.; Friedman, N.; Regev, A. Deciphering Eukaryotic Gene-Regulatory Logic with 100 Million Random Promoters. *Nat. Biotechnol.* **2019**, 1–10. https://doi.org/10.1038/s41587-019-0315-8.
(77) Vaishnav, E. D.; de Boer, C. G.; Molinet, J.; Yassour, M.; Fan, L.; Adiconis, X.; Thompson, D. A.; Levin, J. Z.; Cubillos, F. A.; Regev, A. The Evolution, Evolvability and Engineering of Gene Regulatory DNA. *Nature* **2022**, 1–9. https://doi.org/10.1038/s41586-022-04506-6.
(78) Alipanahi, B.; Delong, A.; Weirauch, M. T.; Frey, B. J. Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* **2015**, *33* (8), 831–838. https://doi.org/10.1038/nbt.3300.
(79) Zhou, J.; Troyanskaya, O. G. Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. *Nat. Methods* **2015**, *12* (10), 931–934. https://doi.org/10.1038/nmeth.3547.
(80) Kelley, D. R.; Snoek, J.; Rinn, J. L. Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks. *Genome Res.* **2016**, *26* (7), 990–999. https://doi.org/10.1101/gr.200535.115.
(81) Hiranuma, N.; Lundberg, S.; Lee, S.-I. DeepATAC: A Deep-Learning Method to Predict Regulatory Factor Binding Activity from ATAC-Seq Signals. *bioRxiv* **2017**, 172767. https://doi.org/10.1101/172767.
(82) Harley, C. B.; Reynolds, R. P. Analysis of E. Coli Promoter Sequences. *Nucleic Acids Res.* **1987**, *15* (5), 2343–2361. https://doi.org/10.1093/nar/15.5.2343.
(83) Vera, J. M.; Ghosh, I. N.; Zhang, Y.; Hebert, A. S.; Coon, J. J.; Landick, R. Genome-Scale Transcription-Translation Mapping Reveals Features of Zymomonas Mobilis Transcription Units and Promoters. *mSystems* **2020**, *5* (4). https://doi.org/10.1128/mSystems.00250-20.
(84) Bailey, T. L.; Elkan, C. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Mach. Learn.* **1995**, *21* (1), 51–80. https://doi.org/10.1007/BF00993379.
(85) Gordon, L.; Chervonenkis, A. Y.; Gammerman, A. J.; Shahmuradov, I. A.; Solovyev, V. V.

Sequence Alignment Kernel for Recognition of Promoter Regions. *Bioinformatics* **2003**, *19* (15), 1964–1971. https://doi.org/10.1093/bioinformatics/btg265.
(86) Shimada, T.; Yamazaki, Y.; Tanaka, K.; Ishihama, A. The Whole Set of Constitutive Promoters Recognized by RNA Polymerase RpoD Holoenzyme of Escherichia Coli. *PLOS ONE* **2014**, *9* (3), e90447. https://doi.org/10.1371/journal.pone.0090447.
(87) Kanhere, A.; Bansal, M. A Novel Method for Prokaryotic Promoter Prediction Based on DNA Stability. *BMC Bioinformatics* **2005**, *6* (1), 1. https://doi.org/10.1186/1471-2105-6-1.
(88) Cassiano, M. H. A.; Silva-Rocha, R. Benchmarking Bacterial Promoter Prediction Tools: Potentialities and Limitations. *mSystems* **2020**, *5* (4). https://doi.org/10.1128/mSystems.00439-20.
(89) Santos-Zavaleta, A.; Salgado, H.; Gama-Castro, S.; Sánchez-Pérez, M.; Gómez-Romero, L.; Ledezma-Tejeida, D.; García-Sotelo, J. S.; Alquicira-Hernández, K.; Muñiz-Rascado, L. J.; Peña-Loredo, P.; Ishida-Gutiérrez, C.; Velázquez-Ramírez, D. A.; Del Moral-Chávez, V.; Bonavides-Martínez, C.; Méndez-Cruz, C.-F.; Galagan, J.; Collado-Vides, J. RegulonDB v 10.5: Tackling Challenges to Unify Classic and High Throughput Knowledge of Gene Regulation in E. Coli K-12. *Nucleic Acids Res.* **2019**, *47* (D1), D212–D220. https://doi.org/10.1093/nar/gky1077.
(90) Hershberg, R.; Bejerano, G.; Santos-Zavaleta, A.; Margalit, H. PromEC: An Updated Database of Escherichia Coli MRNA Promoters with Experimentally Identified Transcriptional Start Sites. *Nucleic Acids Res.* **2001**, *29* (1), 277. https://doi.org/10.1093/nar/29.1.277.
(91) Keseler, I. M.; Mackie, A.; Santos-Zavaleta, A.; Billington, R.; Bonavides-Martínez, C.; Caspi, R.; Fulcher, C.; Gama-Castro, S.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Muñiz-Rascado, L.; Ong, Q.; Paley, S.; Peralta-Gil, M.; Subhraveti, P.; Velázquez-Ramírez, D. A.; Weaver, D.; Collado-Vides, J.; Paulsen, I.; Karp, P. D. The EcoCyc Database: Reflecting New Knowledge about Escherichia Coli K-12. *Nucleic Acids Res.* **2017**, *45* (D1), D543–D550. https://doi.org/10.1093/nar/gkw1003.
(92) Sharma, C. M.; Vogel, J. Differential RNA-Seq: The Approach behind and the Biological Insight Gained. *Curr. Opin. Microbiol.* **2014**, *19*, 97–105. https://doi.org/10.1016/j.mib.2014.06.010.
(93) Green, M. R.; Sambrook, J. Rapid Amplification of Sequences from the 5’ Ends of MRNAs: 5’-RACE. *Cold Spring Harb. Protoc.* **2019**, *2019* (5). https://doi.org/10.1101/pdb.prot095208.
(94) Fu, Y.; Li, Y.; Lidstrom, M. The Oxidative TCA Cycle Operates during Methanotrophic Growth of the Type I Methanotroph Methylomicrobium Buryatense 5GB1. *Metab. Eng.* **2017**, *42*, 43–51. https://doi.org/10.1016/j.ymben.2017.05.003.
(95) He, L.; Fu, Y.; Lidstrom, M. E. Quantifying Methane and Methanol Metabolism of “Methylotuvimicrobium Buryatense” 5GB1C under Substrate Limitation. *mSystems* **2019**, *4* (6). https://doi.org/10.1128/mSystems.00748-19.
(96) Groom, J. D.; Ford, S. M.; Pesesky, M. W.; Lidstrom, M. E. A Mutagenic Screen Identifies a TonB-Dependent Receptor Required for the Lanthanide Metal Switch in the Type I Methanotroph “Methylotuvimicrobium Buryatense” 5GB1C. *J. Bacteriol.* **2019**, *201* (15). https://doi.org/10.1128/JB.00120-19.
(97) *Promoters/Catalog/Anderson - parts.igem.org*. http://parts.igem.org/Promoters/Catalog/Anderson (accessed 2021-03-30).
(98) Gilman, J.; Love, J. Synthetic Promoter Design for New Microbial Chassis. *Biochem. Soc. Trans.* **2016**, *44* (3), 731–737. https://doi.org/10.1042/BST20160042.
(99) Redden, H.; Alper, H. S. The Development and Characterization of Synthetic Minimal Yeast Promoters. *Nat. Commun.* **2015**, *6*, 7810. https://doi.org/10.1038/ncomms8810.
(100) Liu, X.; Brutlag, D. L.; Liu, J. S. BioProspector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes. *Pac. Symp. Biocomput. Pac. Symp.*

Biocomput. **2001**, 127–138.
(101) Osbourn, A. E.; Field, B. Operons. *Cell. Mol. Life Sci. CMLS* **2009**, *66* (23), 3755–3775. https://doi.org/10.1007/s00018-009-0114-3.
(102) Tompa, M.; Li, N.; Bailey, T. L.; Church, G. M.; Moor, B. D.; Eskin, E.; Favorov, A. V.; Frith, M. C.; Fu, Y.; Kent, W. J.; Makeev, V. J.; Mironov, A. A.; Noble, W. S.; Pavesi, G.; Pesole, G.; Régnier, M.; Simonis, N.; Sinha, S.; Thijs, G.; Helden, J. van; Vandenbogaert, M.; Weng, Z.; Workman, C.; Ye, C.; Zhu, Z. Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites. *Nat. Biotechnol.* **2005**, *23* (1), 137–144. https://doi.org/10.1038/nbt1053.
(103) Tareen, A.; Kinney, J. B. Logomaker: Beautiful Sequence Logos in Python. *Bioinformatics* **2020**, *36* (7), 2272–2274. https://doi.org/10.1093/bioinformatics/btz921.
(104) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.
(105) Nair, T. M.; Kulkarni, B. D. On the Consensus Structure within the E. Coli Promoters. *Biophys. Chem.* **1994**, *48* (3), 383–393. https://doi.org/10.1016/0301-4622(93)E0056-B.
(106) Jensen, P. R.; Hammer, K. The Sequence of Spacers between the Consensus Sequences Modulates the Strength of Prokaryotic Promoters. *Appl. Environ. Microbiol.* **1998**, *64* (1), 82–87. https://doi.org/10.1128/AEM.64.1.82-87.1998.
(107) Gaballa, A.; Guariglia-Oropeza, V.; Dürr, F.; Butcher, B. G.; Chen, A. Y.; Chandrangsu, P.; Helmann, J. D. Modulation of Extracytoplasmic Function (ECF) Sigma Factor Promoter Selectivity by Spacer Region Sequence. *Nucleic Acids Res.* **2018**, *46* (1), 134–145. https://doi.org/10.1093/nar/gkx953.
(108) Nicolas, P.; Mäder, U.; Dervyn, E.; Rochat, T.; Leduc, A.; Pigeonneau, N.; Bidnenko, E.; Marchadier, E.; Hoebeke, M.; Aymerich, S.; Becher, D.; Bisicchia, P.; Botella, E.; Delumeau, O.; Doherty, G.; Denham, E. L.; Fogg, M. J.; Fromion, V.; Goelzer, A.; Hansen, A.; Härtig, E.; Harwood, C. R.; Homuth, G.; Jarmer, H.; Jules, M.; Klipp, E.; Chat, L. L.; Lecointe, F.; Lewis, P.; Liebermeister, W.; March, A.; Mars, R. A. T.; Nannapaneni, P.; Noone, D.; Pohl, S.; Rinn, B.; Rügheimer, F.; Sappa, P. K.; Samson, F.; Schaffer, M.; Schwikowski, B.; Steil, L.; Stülke, J.; Wiegert, T.; Devine, K. M.; Wilkinson, A. J.; Dijl, J. M. van; Hecker, M.; Völker, U.; Bessières, P.; Noirot, P. Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus Subtilis. *Science* **2012**, *335* (6072), 1103–1106. https://doi.org/10.1126/science.1206848.
(109) Latif, H.; Lerman, J. A.; Portnoy, V. A.; Tarasova, Y.; Nagarajan, H.; Schrimpe-Rutledge, A. C.; Smith, R. D.; Adkins, J. N.; Lee, D.-H.; Qiu, Y.; Zengler, K. The Genome Organization of Thermotoga Maritima Reflects Its Lifestyle. *PLOS Genet.* **2013**, *9* (4), e1003485. https://doi.org/10.1371/journal.pgen.1003485.
(110) Wilson, E. H.; Groom, J. D.; Sarfatis, M. C.; Ford, S. M.; Lidstrom, M. E.; Beck, D. A. C. A Computational Framework for Identifying Promoter Sequences in Nonmodel Organisms Using RNA-Seq Data Sets. *ACS Synth. Biol.* **2021**. https://doi.org/10.1021/acssynbio.1c00017.
(111) Wu, G.; Yan, Q.; Jones, J. A.; Tang, Y. J.; Fong, S. S.; Koffas, M. A. G. Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. *Trends Biotechnol.* **2016**, *34* (8), 652–664. https://doi.org/10.1016/j.tibtech.2016.02.010.
(112) Anesiadis, N.; Cluett, W. R.; Mahadevan, R. Dynamic Metabolic Engineering for Increasing Bioprocess Productivity. *Metab. Eng.* **2008**, *10* (5), 255–266. https://doi.org/10.1016/j.ymben.2008.06.004.
(113) Brockman, I. M.; Prather, K. L. J. Dynamic Metabolic Engineering: New Strategies for Developing Responsive Cell Factories. *Biotechnol. J.* **2015**, *10* (9), 1360–1369. https://doi.org/10.1002/biot.201400422.

(114) Gardner, T. S.; Cantor, C. R.; Collins, J. J. Construction of a Genetic Toggle Switch in Escherichia Coli. *Nature* **2000**, *403* (6767), 339–342. https://doi.org/10.1038/35002131.
(115) Han, T.; Chen, Q.; Liu, H. Engineered Photoactivatable Genetic Switches Based on the Bacterium Phage T7 RNA Polymerase. *ACS Synth. Biol.* **2017**, *6* (2), 357–366. https://doi.org/10.1021/acssynbio.6b00248.
(116) Research, I. of M. (US) C. on M. N. *The Energy Costs of Protein Metabolism: Lean and Mean on Uncle Sam's Team*; National Academies Press (US), 1999.
(117) Endalur Gopinarayanan, V.; Nair, N. U. A Semi-Synthetic Regulon Enables Rapid Growth of Yeast on Xylose. *Nat. Commun.* **2018**, *9* (1), 1233. https://doi.org/10.1038/s41467-018-03645-7.
(118) Latchman, D. S. Transcription Factors: An Overview. *Int. J. Biochem. Cell Biol.* **1997**, *29* (12), 1305–1312. https://doi.org/10.1016/S1357-2725(97)00085-X.
(119) Segal, E.; Shapira, M.; Regev, A.; Pe'er, D.; Botstein, D.; Koller, D.; Friedman, N. Module Networks: Identifying Regulatory Modules and Their Condition-Specific Regulators from Gene Expression Data. *Nat. Genet.* **2003**, *34* (2), 166–176. https://doi.org/10.1038/ng1165.
(120) Baliga, N. S.; Björkegren, J. L. M.; Boeke, J. D.; Boutros, M.; Crawford, N. P. S.; Dudley, A. M.; Farber, C. R.; Jones, A.; Levey, A. I.; Lusis, A. J.; Mak, H. C.; Nadeau, J. H.; Noyes, M. B.; Petretto, E.; Seyfried, N. T.; Steinmetz, L. M.; Vonesch, S. C. The State of Systems Genetics in 2017. *Cell Syst.* **2017**, *4* (1), 7–15. https://doi.org/10.1016/j.cels.2017.01.005.
(121) Reiss, D. J.; Baliga, N. S.; Bonneau, R. Integrated Biclustering of Heterogeneous Genome-Wide Datasets for the Inference of Global Regulatory Networks. *BMC Bioinformatics* **2006**, *7* (1), 280. https://doi.org/10.1186/1471-2105-7-280.
(122) Reiss, D. J.; Plaisier, C. L.; Wu, W.-J.; Baliga, N. S. CMonkey2: Automated, Systematic, Integrated Detection of Co-Regulated Gene Modules for Any Organism. *Nucleic Acids Res.* **2015**, *43* (13), e87–e87. https://doi.org/10.1093/nar/gkv300.
(123) Herring, C. D.; Raffaelle, M.; Allen, T. E.; Kanin, E. I.; Landick, R.; Ansari, A. Z.; Palsson, B. Ø. Immobilization of Escherichia Coli RNA Polymerase and Location of Binding Sites by Use of Chromatin Immunoprecipitation and Microarrays. *J. Bacteriol.* **2005**, *187* (17), 6166–6174. https://doi.org/10.1128/JB.187.17.6166-6174.2005.
(124) Stormo, G. D. DNA Binding Sites: Representation and Discovery. *Bioinformatics* **2000**, *16* (1), 16–23. https://doi.org/10.1093/bioinformatics/16.1.16.
(125) Cardon, L. R.; Stormo, G. D. Expectation Maximization Algorithm for Identifying Protein-Binding Sites with Variable Lengths from Unaligned DNA Fragments. *J. Mol. Biol.* **1992**, *223* (1), 159–170. https://doi.org/10.1016/0022-2836(92)90723-W.
(126) Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. MEME Suite: Tools for Motif Discovery and Searching. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W202–W208. https://doi.org/10.1093/nar/gkp335.
(127) Huang, W.; Umbach, D. M.; Ohler, U.; Li, L. Optimized Mixed Markov Models for Motif Identification. *BMC Bioinformatics* **2006**, *7* (1), 279. https://doi.org/10.1186/1471-2105-7-279.
(128) Comon, P. Independent Component Analysis, A New Concept? *Signal Process.* **1994**, *36* (3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9.
(129) Kong, W.; Vanderburg, C. R.; Gunshin, H.; Rogers, J. T.; Huang, X. A Review of Independent Component Analysis Application to Microarray Gene Expression Data. *BioTechniques* **2008**, *45* (5), 501–520. https://doi.org/10.2144/000112950.
(130) Rychel, K.; Decker, K.; Sastry, A. V.; Phaneuf, P. V.; Poudel, S.; Palsson, B. O. IModulonDB: A Knowledgebase of Microbial Transcriptional Regulation Derived from Machine Learning. *Nucleic Acids Res.* **2021**, *49* (D1), D112–D120. https://doi.org/10.1093/nar/gkaa810.
(131) Saelens, W.; Cannoodt, R.; Saeys, Y. A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data. *Nat. Commun.* **2018**, *9* (1), 1090.

https://doi.org/10.1038/s41467-018-03424-4.
(132) Gilman, A. Development of a Promising Methanotrophic Bacterium as an Industrial Biocatalyst. Thesis, University of Washington, Seattle, WA, 2017. https://digital.lib.washington.edu:443/researchworks/handle/1773/39973 (accessed 2020-12-15).
(133) Kelley, D. R.; Reshef, Y. A.; Bileschi, M.; Belanger, D.; McLean, C. Y.; Snoek, J. Sequential Regulatory Activity Prediction across Chromosomes with Convolutional Neural Networks. *Genome Res.* **2018**, *28* (5), 739–750. https://doi.org/10.1101/gr.227819.117.
(134) Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; Kelley, D. R. Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions. *Nat. Methods* **2021**, *18* (10), 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.
(135) Buenrostro, J. D.; Giresi, P. G.; Zaba, L. C.; Chang, H. Y.; Greenleaf, W. J. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nat. Methods* **2013**, *10* (12), 1213–1218. https://doi.org/10.1038/nmeth.2688.
(136) Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep Learning for Computational Biology. *Mol. Syst. Biol.* **2016**, *12* (7), 878. https://doi.org/10.15252/msb.20156651.
(137) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2021**, 1–16. https://doi.org/10.1038/s41580-021-00407-0.
(138) Camacho, D. M.; Collins, K. M.; Powers, R. K.; Costello, J. C.; Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell* **2018**, *173* (7), 1581–1592. https://doi.org/10.1016/j.cell.2018.05.015.
(139) Lee, H.; Grosse, R.; Ranganath, R.; Ng, A. Y. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*; ACM: Montreal Quebec Canada, 2009; pp 609–616. https://doi.org/10.1145/1553374.1553453.
(140) Park, Y.; Kellis, M. Deep Learning for Regulatory Genomics. *Nat. Biotechnol.* **2015**, *33* (8), 825–826. https://doi.org/10.1038/nbt.3313.
(141) Koo, P. K.; Ploenzke, M. Deep Learning for Inferring Transcription Factor Binding Sites. *Curr. Opin. Syst. Biol.* **2020**, *19*, 16–23. https://doi.org/10.1016/j.coisb.2020.04.001.
(142) Quang, D.; Xie, X. DanQ: A Hybrid Convolutional and Recurrent Deep Neural Network for Quantifying the Function of DNA Sequences. *Nucleic Acids Res.* **2016**, *44* (11), e107–e107. https://doi.org/10.1093/nar/gkw226.
(143) Jurtz, V. I.; Johansen, A. R.; Nielsen, M.; Almagro Armenteros, J. J.; Nielsen, H.; Sønderby, C. K.; Winther, O.; Sønderby, S. K. An Introduction to Deep Learning on Biological Sequence Data: Examples and Solutions. *Bioinformatics* **2017**, *33* (22), 3685–3690. https://doi.org/10.1093/bioinformatics/btx531.
(144) Agarwal, V.; Shendure, J. Predicting MRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **2020**, *31* (7), 107663. https://doi.org/10.1016/j.celrep.2020.107663.
(145) Hiranuma, N.; Park, H.; Baek, M.; Anishchenko, I.; Dauparas, J.; Baker, D. Improved Protein Structure Refinement Guided by Deep Learning Based Accuracy Estimation. *Nat. Commun.* **2021**, *12* (1), 1340. https://doi.org/10.1038/s41467-021-21511-x.
(146) *Basics | Stars*. NASA Universe Exploration. https://universe.nasa.gov/stars/basics (accessed 2023-04-22).
(147) Litterman, A. J.; Kageyama, R.; Tonqueze, O. L.; Zhao, W.; Gagnon, J. D.; Goodarzi, H.; Erle, D. J.; Ansel, K. M. A Massively Parallel 3′ UTR Reporter Assay Reveals Relationships between Nucleotide Content, Sequence Conservation, and MRNA Destabilization. *Genome Res.* **2019**. https://doi.org/10.1101/gr.242552.118.

(148) de Avila e Silva, S.; Echeverrigaray, S.; Gerhardt, G. J. L. BacPP: Bacterial Promoter Prediction—A Tool for Accurate Sigma-Factor Specific Assignment in Enterobacteria. *J. Theor. Biol.* **2011**, *287*, 92–99. https://doi.org/10.1016/j.jtbi.2011.07.017.
(149) Shahmuradov, I. A.; Mohamad Razali, R.; Bougouffa, S.; Radovanovic, A.; Bajic, V. B. BTSSfinder: A Novel Tool for the Prediction of Promoters in Cyanobacteria and Escherichia Coli. *Bioinforma. Oxf. Engl.* **2017**, *33* (3), 334–340. https://doi.org/10.1093/bioinformatics/btw629.
(150) Umarov, R. K.; Solovyev, V. V. Recognition of Prokaryotic and Eukaryotic Promoters Using Convolutional Deep Learning Neural Networks. *PLOS ONE* **2017**, *12* (2), e0171410. https://doi.org/10.1371/journal.pone.0171410.
(151) Rhodius, V. A.; Mutalik, V. K. Predicting Strength and Function for Promoters of the Escherichia Coli Alternative Sigma Factor, SigmaE. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (7), 2854–2859. https://doi.org/10.1073/pnas.0915066107.
(152) Bharanikumar, R.; Premkumar, K. A. R.; Palaniappan, A. PromoterPredict: Sequence-Based Modelling of Escherichia Coli Σ70 Promoter Strength Yields Logarithmic Dependence between Promoter Strength and Sequence. *PeerJ* **2018**, *6*. https://doi.org/10.7717/peerj.5862.
(153) Tayara, H.; Tahir, M.; Chong, K. T. Identification of Prokaryotic Promoters and Their Strength by Integrating Heterogeneous Features. *Genomics* **2020**, *112* (2), 1396–1403. https://doi.org/10.1016/j.ygeno.2019.08.009.
(154) Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; Hanna, A. AI and the Everything in the Whole Wide World Benchmark. *Proc. NeurIPS 2020 Workshop ML Retrosp. Surv. Meta-Anal. ML-RSA* **2020**, 20.
(155) Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; Zhou, Y. *Deep Learning Scaling is Predictable, Empirically*. arXiv.org. https://arxiv.org/abs/1712.00409v1 (accessed 2023-05-11).
(156) Shrikumar, A.; Tian, K.; Avsec, Ž.; Shcherbina, A.; Banerjee, A.; Sharmin, M.; Nair, S.; Kundaje, A. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) Version 0.5.6.5. **2018**.
(157) Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*; ICML'17; JMLR.org: Sydney, NSW, Australia, 2017; pp 3145–3153.
(158) Chen, H.; Lundberg, S.; Lee, S.-I. Explaining Models by Propagating Shapley Values of Local Components. *ArXiv191111888 Cs Stat* **2019**.
(159) Linder, J.; La Fleur, A.; Chen, Z.; Ljubetič, A.; Baker, D.; Kannan, S.; Seelig, G. Interpreting Neural Networks for Biological Sequences by Learning Stochastic Masks. *Nat. Mach. Intell.* **2022**, *4* (1), 41–54. https://doi.org/10.1038/s42256-021-00428-6.
(160) Johnson, J. M.; Khoshgoftaar, T. M. Survey on Deep Learning with Class Imbalance. *J. Big Data* **2019**, *6* (1), 27. https://doi.org/10.1186/s40537-019-0192-5.
(161) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
(162) Thomason, M. K.; Bischler, T.; Eisenbart, S. K.; Förstner, K. U.; Zhang, A.; Herbig, A.; Nieselt, K.; Sharma, C. M.; Storz, G. Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia Coli. *J. Bacteriol.* **2015**, *197* (1), 18–28. https://doi.org/10.1128/JB.02096-14.
(163) Urtecho, G.; Insigne, K. D.; Tripp, A. D.; Brinck, M.; Lubock, N. B.; Kim, H.; Chan, T.; Kosuri, S. Genome-Wide Functional Characterization of Escherichia Coli Promoters and Regulatory Elements Responsible for Their Function. bioRxiv January 6, 2020, p 2020.01.04.894907. https://doi.org/10.1101/2020.01.04.894907.

(164) Wang, Y.; Wang, H.; Wei, L.; Li, S.; Liu, L.; Wang, X. Synthetic Promoter Design in Escherichia Coli Based on a Deep Generative Network. *Nucleic Acids Res.* **2020**, *48* (12), 6403–6412. https://doi.org/10.1093/nar/gkaa325.
(165) Zrimec, J.; Börlin, C. S.; Buric, F.; Muhammad, A. S.; Chen, R.; Siewers, V.; Verendel, V.; Nielsen, J.; Töpel, M.; Zelezniak, A. Deep Learning Suggests That Gene Expression Is Encoded in All Parts of a Co-Evolving Interacting Gene Regulatory Structure. *Nat. Commun.* **2020**, *11* (1), 6141. https://doi.org/10.1038/s41467-020-19921-4.
(166) Park, S.; Koh, Y.; Jeon, H.; Kim, H.; Yeo, Y.; Kang, J. Enhancing the Interpretability of Transcription Factor Binding Site Prediction Using Attention Mechanism. *Sci. Rep.* **2020**, *10* (1), 13413. https://doi.org/10.1038/s41598-020-70218-4.
(167) Fornes, O.; Castro-Mondragon, J. A.; Khan, A.; van der Lee, R.; Zhang, X.; Richmond, P. A.; Modi, B. P.; Correard, S.; Gheorghe, M.; Baranašić, D.; Santana-Garcia, W.; Tan, G.; Chèneby, J.; Ballester, B.; Parcy, F.; Sandelin, A.; Lenhard, B.; Wasserman, W. W.; Mathelier, A. JASPAR 2020: Update of the Open-Access Database of Transcription Factor Binding Profiles. *Nucleic Acids Res.* **2020**, *48* (D1), D87–D92. https://doi.org/10.1093/nar/gkz1001.
(168) Rogozin, I. B.; Makarova, K. S.; Natale, D. A.; Spiridonov, A. N.; Tatusov, R. L.; Wolf, Y. I.; Yin, J.; Koonin, E. V. Congruent Evolution of Different Classes of Non-Coding DNA in Prokaryotic Genomes. *Nucleic Acids Res.* **2002**, *30* (19), 4264–4271.
(169) Mendoza-Vargas, A.; Olvera, L.; Olvera, M.; Grande, R.; Vega-Alvarado, L.; Taboada, B.; Jimenez-Jacinto, V.; Salgado, H.; Juárez, K.; Contreras-Moreira, B.; Huerta, A. M.; Collado-Vides, J.; Morett, E. Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. Coli. *PLoS ONE* **2009**, *4* (10), e7526. https://doi.org/10.1371/journal.pone.0007526.
(170) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A Survey of Transfer Learning. *J. Big Data* **2016**, *3* (1), 9. https://doi.org/10.1186/s40537-016-0043-6.
(171) Kelley, D. R. Cross-Species Regulatory Sequence Activity Prediction. *PLOS Comput. Biol.* **2020**, *16* (7), e1008050. https://doi.org/10.1371/journal.pcbi.1008050.
(172) Reddy, A. J.; Herschl, M. H.; Kolli, S.; Lu, A. X.; Geng, X.; Kumar, A.; Hsu, P. D.; Levine, S.; Ioannidis, N. M. *Pretraining Strategies for Effective Promoter-Driven Gene Expression Prediction*; preprint; Genomics, 2023. https://doi.org/10.1101/2023.02.24.529941.
(173) Novakovsky, G.; Saraswat, M.; Fornes, O.; Mostafavi, S.; Wasserman, W. W. Biologically Relevant Transfer Learning Improves Transcription Factor Binding Prediction. *Genome Biol.* **2021**, *22* (1), 280. https://doi.org/10.1186/s13059-021-02499-5.
(174) Clauwaert, J.; Waegeman, W. Novel Transformer Networks for Improved Sequence Labeling in Genomics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, 1–1. https://doi.org/10.1109/TCBB.2020.3035021.
(175) Clauwaert, J.; Menschaert, G.; Waegeman, W. Explainability in Transformer Models for Functional Genomics. *Brief. Bioinform.* **2021**, No. bbab060. https://doi.org/10.1093/bib/bbab060.
(176) Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R. V. DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome. *Bioinformatics* **2021**, *37* (15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083.
(177) Wilson, E.; Lidstrom, M.; Beck, D. A Multi-Task Learning Approach to Enhance Sustainable Biomolecule Production in Engineered Microorganisms. In *Climate Change AI*; Climate Change AI, 2021.

Onward and upward…