UNIVERSITY OF CALIFORNIA SAN DIEGO

Integrating Multi-Omics Knowledge with Metabolic Modeling: Enabling Interoperability Across Computational Biology Tools

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioengineering

by

Arjun Patel

Committee in charge:

Professor Bernhard Ø Palsson, Chair
Professor David J Gonzalez
Professor Nathan E Lewis
Professor Benjamin L Smarr

2025

The Dissertation of Arjun Patel is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

Dedication

To my family, friends, best friend/wife, and cat. I couldn't have done this without you all.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my advisor, Professor Bernhard Palsson, for his guidance, creativity, and vision throughout this work. Every meeting with him was a reminder of how deeply he thinks about biology and systems, his questions always pushed me toward clarity and insight, shaping how I approach research and scientific reasoning.

I am deeply grateful to Dr. Daniel Zielinski, whose mentorship was instrumental in my development as a scientist. His depth of knowledge, patience, and encouragement made him one of the best mentors I could have asked for.

I also want to thank Dr. Amitesh Anand and Dr. Ke Chen, who introduced me to ME modeling and the integration of multi-omics data early in my Ph.D. Their projects and generosity with their time accelerated my learning and helped define the direction of my work.

I am grateful to all the members of the Systems Biology Research Group and colleagues at the Center for Biosustainability at DTU Biosustain in Denmark. Their discussions, feedback, and collaboration made this an intellectually rich and supportive environment.

To my family, thank you for instilling in me the values of hard work, curiosity, and perseverance. To my wife, thank you for being my foundation through every high and low, your unwavering support and patience made this possible. And to my cat, thank you for the constant entertainment and quiet company during long days of writing and coding.

Finally, I gratefully acknowledge the support of the Novo Nordisk Foundation and the National Institutes of Health, whose funding made this work possible.

Chapter 2, in part, is a reprint of the material as it appears in *Nature Communications*. **- Patel, Arjun**; McGrosso, Dominic; Hefner, Ying; Campeau, Anaamika; Sastry, Anand V; Maurya, Svetlana; Rychel, Kevin; Gonzalez, David J; Palsson, Bernhard O (2024). *Proteome*

allocation is linked to transcriptional regulation through a modularized transcriptome. Nature Communications, 15(1), 5234. The dissertation author was the primary author.

Chapter 3, in part, is a reprint of material as it appears in *Nature Communications,* and in part, is a reprint of material submitted to *iScience*.

- Anand, Amitesh; **Patel, Arjun**; Chen, Ke; Olson, Connor A; Phaneuf, Patrick V; Lamoureux, Cameron; Hefner, Ying; Szubin, Richard; Feist, Adam M; Palsson, Bernhard O (2022). *Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity.* Nature Communications, 13(1), 3682. The dissertation author was the second author of this paper.

- Patel, Arjun; Banwani, Neha; Mink, Raegan; Prabhakaran, Divya M; Khairnar, Snehal V; Feist, Adam M; Palsson, Bernhard O; Anand, Amitesh (2025). *Aerobicity stimulon in Escherichia coli revealed using multi-scale computational systems biology of adapted respiratory variants.* bioRxiv, 2025.03.13.642450. The dissertation author was the primary author of this paper.

Chapter 4, in part, is a reprint of material as it appears in *mSystems, Cell Reports, mSystems, Cell,* and *Cell Reports.*

- Dalldorf, Christopher; Rychel, Kevin; Szubin, Richard; Hefner, Ying; **Patel, Arjun**; Zielinski, Daniel C; Palsson, Bernhard O (2024). *The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions.* mSystems, 9(7), e00305-24. The dissertation author was the fifth author of this material.

- Rychel, Kevin; Tan, Justin; **Patel, Arjun**; Lamoureux, Cameron; Hefner, Ying; Szubin, Richard; Johnsen, Josefin; Mohamed, Elsayed Tharwat Tolba; Phaneuf, Patrick V; Anand, Amitesh; Olson, Connor A; Park, Joon Ho; Sastry, Anand V; Yang, Laurence; Feist, Adam M;

Palsson, Bernhard O (2023). *Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance.* Cell Reports, 42(9). The dissertation author was the third author of this material.

- Beulig, Felix; Bafna-Rührer, Jonas; Jensen, Peter E; Kim, SH; **Patel, Arjun**; Kandasamy, Vijayalakshmi; Steffen, CS; Decker, Katherine; Zielinski, DC; Yang, Laurence; Ozdemir, Emre; Sudarsan, Suresh; Palsson, Bernhard O (2025). *Trade-off between resistance and persistence in high cell density Escherichia coli cultures.* mSystems, 10(7), e0032325. The dissertation author was the fifth author of this material.

- Kundu, Biki Bapi; Krishnan, Jayanth; Szubin, Richard; **Patel, Arjun**; Palsson, Bernhard O; Zielinski, Daniel C; Ajo-Franklin, Caroline M (2025). *Extracellular respiration is a latent energy metabolism in Escherichia coli.* Cell, 188(11), 2907–2924.e23. The dissertation author was the fourth author of this material.

- Shin, Jongoh; **Patel, Arjun**; Lou Xuwen A; Catoiu, Edward; Krishnan, Jayanth; Hefner, Ying; Szubin, Richard; Sung, Jaemin; Francis Son, Hyeoncheol, Zielinski, Daniel C; Palsson, Bernhard O (2025). *Escherichia coli chooses substrates using a multi-layered hierarchical regulatory network*. The dissertation author was the second author.

- Anand, Amitesh; Olson, Connor A; Sastry, Anand V; **Patel, Arjun**; Szubin, Richard; Yang, Laurence; Feist, Adam M; Palsson, Bernhard O (2021). *Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications.* Cell Reports, 35(1), 108961. The dissertation author was the fourth author of this material.

Chapter 5 in part, is currently being prepared for submission for publication.

- Patel, Arjun; Catoiu, Edward A; Palsson, Bernhard O (2025). *COBRAme.org: A web application for solving and constraining metabolism and expression models.* The dissertation author was the primary author.

Vita

2019 Bachelor of Science in Biomedical Engineering, Georgia Institute of Technology

2025 Doctor of Philosophy in Bioengineering, University of California San Diego

Publications

Rychel, Kevin; Chen, Ke; Catoiu, Edward A; Olson, Connor A; Sandberg, Troy E; Gao, Ye; Xu, Sibei; Hefner, Ying; Szubin, Richard; **Patel, Arjun**; Feist, Adam M; Palsson, Bernhard O (2024). *Laboratory evolution reveals transcriptional mechanisms underlying thermal adaptation of Escherichia coli.* **Genome Biol Evol**, 2025 Sep 30;17(10):evaf171.

Kundu, Biki Bapi; Krishnan, Jayanth; Szubin, Richard; **Patel, Arjun**; Palsson, Bernhard O; Zielinski, Daniel C; Ajo-Franklin, Caroline M (2025). *Extracellular respiration is a latent energy metabolism in Escherichia coli.* **Cell**, 188(11), 2907–2924.e23.

Catoiu, Edward A; Krishnan, Jayanth; Li, Gaoyuan; Lou, Xuwen A; Rychel, Kevin; Yuan, Yuan; Bajpe, Heera; **Patel, Arjun**; Choe, Donghui; Shin, Jongoh; Burrows, Joshua; Phaneuf, Patrick V; Zielinski, Daniel C; Palsson, Bernhard O (2025). *iModulonDB 2.0: dynamic tools to facilitate knowledge-mining and user-enabled analyses of curated transcriptomic datasets.* **Nucleic Acids Research**, 53(D1), D99–D106.

Patel, Arjun; Banwani, Neha; Mink, Raegan; Prabhakaran, Divya M; Khairnar, Snehal V; Feist, Adam M; Palsson, Bernhard O; Anand, Amitesh (2025). *Aerobicity stimulon in Escherichia coli revealed using multi-scale computational systems biology of adapted respiratory variants.* **bioRxiv**, 2025.03.13.642450.

Beulig, Felix; Bafna-Rührer, Jonas; Jensen, Peter E; Kim, SH; **Patel, Arjun**; Kandasamy, Vijayalakshmi; Steffen, CS; Decker, Katherine; Zielinski, DC; Yang, Laurence; Ozdemir, Emre; Sudarsan, Suresh; Palsson, Bernhard O (2025). *Trade-off between resistance and persistence in high cell density Escherichia coli cultures.* **mSystems**, 10(7), e0032325.

Patel, Arjun; McGrosso, Dominic; Hefner, Ying; Campeau, Anaamika; Sastry, Anand V; Maurya, Svetlana; Rychel, Kevin; Gonzalez, David J; Palsson, Bernhard O (2024). *Proteome allocation is linked to transcriptional regulation through a modularized transcriptome.* **Nature Communications**, 15(1), 5234.

Dalldorf, Christopher; Rychel, Kevin; Szubin, Richard; Hefner, Ying; **Patel, Arjun**; Zielinski, Daniel C; Palsson, Bernhard O (2024). *The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions.* **mSystems**, 9(7), e00305-24.

Rychel, Kevin; Tan, Justin; **Patel, Arjun**; Lamoureux, Cameron; Hefner, Ying; Szubin, Richard; Johnsen, Josefin; Mohamed, Elsayed Tharwat Tolba; Phaneuf, Patrick V; Anand, Amitesh; Olson, Connor A; Park, Joon Ho; Sastry, Anand V; Yang, Laurence; Feist, Adam M; Palsson, Bernhard O (2023). *Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance.* **Cell Reports**, 42(9):113105.

Goel, Nikita; Srivastav, Stuti; **Patel, Arjun**; Shirsath, Akshay; Panda, Tushar Ranjan; Patra, Malay; Feist, Adam M; Anand, Amitesh (2023). *TCA cycle tailoring facilitates optimal growth of proton-pumping NADH dehydrogenase-dependent Escherichia coli.* **Microbiology Spectrum**, 11(6), e02225-23.

Anand, Amitesh; **Patel, Arjun**; Chen, Ke; Olson, Connor A; Phaneuf, Patrick V; Lamoureux, Cameron; Hefner, Ying; Szubin, Richard; Feist, Adam M; Palsson, Bernhard O (2022). *Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity.* **Nature Communications**, 13(1), 3682.

Anand, Amitesh; Olson, Connor A; Sastry, Anand V; **Patel, Arjun**; Szubin, Richard; Yang, Laurence; Feist, Adam M; Palsson, Bernhard O (2021). *Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications.* **Cell Reports**, 35(1), 108961.

Zielinski, Daniel Craig; **Patel, Arjun**; Palsson, Bernhard O (2020). *The expanding computational toolbox for engineering microbial phenotypes at the genome scale.* **Microorganisms**, 8(12), 2050.

Imbach, Kathleen; **Patel, Arjun**; Levine, Aaron D (2018). *Ethical considerations in the translation of CAR-T cell therapies.* **Immuno-oncology Insights**.

Abstract of the Dissertation

Integrating Multi-Omics Knowledge with Metabolic Modeling: Enabling Interoperability Across Computational Biology Tools

by

Arjun Patel

Doctor of Philosophy in Bioengineering

University of California San Diego, 2025

Professor Bernhard Ø Palsson, Chair

Cells continually rebalance limited transcriptional and proteomic resources to survive and grow across environments, but the analytical (omics-scale) and mechanistic (model-based) tools that explain this balance often operate in isolation. This dissertation bridges that divide by integrating knowledge-enriched multi-omics data analytics with proteome-constrained genome-scale modeling to enable interoperable discovery and prediction. First, we show that the bacterial proteome is modularized in a manner that mirrors transcriptomic modules: matched proteome

and transcriptome module components share gene content and condition-dependent activities, revealing how transcriptional and post-translational regulation shape proteome composition. Crucially, these modules enable inference of absolute proteome allocation directly from transcriptomic signals, quantifying resource trade-offs without requiring proteomics in every condition. Second, we integrate experimental data with Metabolism and Expression (ME) models to dissect respiratory plasticity in engineered and evolved electron transport system variants, identifying an "aerobicity stimulon" that captures coordinated shifts in energy metabolism and proteome allocation across aerobic states. Third, we apply this interoperable workflow across diverse case studies, oxidative stress, high-cell-density physiology, extracellular respiration, and hierarchical substrate choice, to quantify condition-specific proteome reallocations and connect mutations, regulation, and phenotype. Finally, we encapsulate these methods in COBRAme.org, the first fully web-based platform for proteome-constrained ME modeling, which separates a lightweight user interface from scalable cloud workers to eliminate software barriers and standardize reproducible analyses. Collectively, the results demonstrate that experimental data and proteome sector derived constraints measurably improve mechanistic predictions, advancing a practical path from big-data signal discovery to explanatory, predictive models. This work provides both conceptual and software infrastructure for interoperable, knowledge-enriched modeling that connects sequencing-scale measurements to genome-scale physiology.

Chapter 1. Introduction

Understanding how biological systems allocate proteomic and energetic resources across environments is a central challenge in systems biology. Recent advances in multi-omics technologies have provided unprecedented visibility into the transcriptional and translational states of cells, yet analytical and modeling frameworks often remain siloed. The goal of this dissertation, to enable the interoperability of multi-omics knowledge-enriched data analytics and tools, arises from the need to connect data-driven biological insight with mechanistic, constraint-based modeling frameworks that can explain and predict cellular behavior.

This introduction is divided into five sections. Sections 1 and 2 describe the conceptual foundations of constraint-based modeling and its expansion to include gene expression. Sections 3 and 4 outline how the big data revolution and data-driven decomposition methods, particularly Independent Component Analysis (ICA) and iModulons, transformed the ability to extract structure from transcriptomics and proteomics. Section 5 then connects these analytical developments to the mechanistic modeling framework, demonstrating how integrating omics data into genome-scale models can refine biological predictions and enable new discoveries.

1.1 Constraint-Based Modeling and Its Roots

Constraint-based modeling (CBM) emerged from the realization that even in the absence of detailed kinetic parameters, the feasible behaviors of a metabolic network are strongly constrained by stoichiometry, thermodynamics, and mass balance. In the late 1990s and early 2000s, this led to the formulation of Flux Balance Analysis (FBA), where cellular metabolism is represented as a system of linear equations defining the steady-state mass balance for every metabolite, and feasible

flux distributions are obtained by solving a linear optimization problem that maximizes an objective function such as growth rate (1).

This simple yet powerful approach made it possible to simulate genome-scale metabolism by leveraging annotated reaction networks rather than empirical rate laws. The first generation of genome-scale metabolic reconstructions, such as *iJR904* and *iAF1260* for *E. coli*, demonstrated that purely stoichiometric models could accurately predict growth yields, essential genes, and substrate utilization patterns across diverse environments (2–4).

A key conceptual strength of CBM is its mechanistic transparency: every predicted phenotype results from a system of mass-balanced reactions subject to physicochemical and environmental constraints. This structure allows researchers to reason about resource allocation, trade-offs, and environmental adaptation quantitatively rather than qualitatively. However, classic FBA treated enzymes as implicitly available catalysts with infinite capacity, ignoring the energetic and proteomic costs of gene expression. This omission motivated the next generation of models that would explicitly couple metabolism to macromolecular expression.

1.2 The Development of Gene Expression Models

As systems biology matured, it became clear that the cell's proteome budget, the fraction of total protein mass allocated to different processes, plays a decisive role in shaping metabolic states. Traditional metabolic models could describe the distribution of fluxes but not the cost of sustaining those fluxes through enzyme synthesis.

To address this limitation, the Metabolism and Expression (ME) modeling framework was developed (5,6). ME models extend FBA by explicitly representing transcription, translation, and complex formation for every catalytic gene product. Each metabolic reaction is linked to the

biosynthetic machinery required to produce its enzyme, introducing coupling constraints that ensure the rate of enzyme synthesis matches the catalytic demand imposed by the metabolic fluxes.

This coupling bridges molecular biology and systems physiology: RNA polymerase and ribosome usage now compete with metabolic enzymes for finite cellular resources, allowing ME models to predict how growth rate, gene expression, and proteome allocation co-vary across conditions. Such explicit accounting transformed genome-scale modeling from a flux-only description into a proteome-constrained model that can explain trade-offs between growth and stress readiness, specialization and generalism, and metabolic efficiency versus robustness.

These mechanistic links paved the way for quantitative integration of omics measurements, especially transcriptomics and proteomics, into constraint-based frameworks. Yet, before such integration could be fully realized, the life sciences underwent a data revolution that changed how biological systems are analyzed.

1.3 The Big Data Revolution

The sequencing and omics revolutions transformed biology from a hypothesis-driven field into a data-saturated discipline. Advances in next-generation sequencing, high-throughput mass spectrometry, and automated culturing produced terabytes of transcriptomic and proteomic data across thousands of conditions (7). At the same time, the rise of FAIR data principles (Findable, Accessible, Interoperable, and Reusable) led to the creation of open databases such as NCBI SRA, EcoCyc, RegulonDB, UniProt, and iModulonDB, enabling unprecedented data reuse and comparative analyses (8–12).

This explosion of public datasets demanded new analytical frameworks capable of extracting biologically meaningful structure from thousands of gene expression profiles. Traditional differential expression analyses, designed for pairwise comparisons, were ill suited for

such high-dimensional data. What emerged instead were unsupervised learning methods that could decompose transcriptomic variation into independent signals, each corresponding to regulatory or functional modules.

In parallel, omics-driven insights began influencing the design of constraint-based models. The availability of thousands of growth condition specific datasets made it possible to contextualize model predictions and test hypotheses about condition dependent regulation, proteome investment, and trade-offs between growth and survival. However, making these datasets interoperable with mechanistic models required new ways of summarizing them, bridging the gap between descriptive data analytics and predictive modeling.

1.4 Independent Component Analysis and iModulons in the Era of Big Data

Independent Component Analysis (ICA) provided one of the first scalable methods for uncovering latent regulatory structure within large transcriptomic compendia. By decomposing the expression matrix into statistically independent components, ICA identifies groups of co-regulated genes, called iModulons, whose activities vary independently across conditions (13).

Each iModulon often corresponds to the activity of a known transcription factor, sigma factor, or global stress response, making them knowledge-enriched analytical units. iModulons therefore bridge data-driven and curated biological understanding: they are derived from statistical decomposition yet interpretable through existing regulatory annotations.

This modular view of transcription dramatically simplifies the analysis of big data compendia. Instead of examining thousands of individual genes, researchers can interpret a smaller set of iModulon activities that capture most of the biologically relevant variance (14). These activities reveal patterns of regulation, such as trade-offs between translation and stress response, that are conserved across environments and even species.

From a modeling standpoint, iModulons represent quantitative summaries that can serve as direct constraints or priors for mechanistic frameworks. For example, a highly active RpoS iModulon could indicate an upregulated stress proteome fraction, which can be translated into limits on the allocation of protein synthesis machinery in an ME model. Conversely, translation related iModulons correspond to growth associated expression sectors that determine the proteomic cost of high growth rates.

By distilling high-dimensional omics into interpretable, biologically grounded features, ICA and iModulons make it possible to integrate data analytics with mechanistic modeling, a central theme of this dissertation.

1.5 Integrating Experimental and Big Data Insights to Refine Model Predictions

The convergence of omics analytics and mechanistic modeling allows the field to move beyond describing correlations to explaining their mechanistic origins. Two key studies, O'Brien et al. 2016 and Yang et al. 2016, demonstrated how integrating proteomic data into ME models can quantitatively improve predictions of growth rate, metabolic fluxes, and proteome composition (15,16).

O'Brien et al. quantified the cost of unused protein expression, showing that roughly half of the *E. coli* proteome mass is unutilized in certain environments, and that this unused fraction imposes measurable fitness costs (15). They formalized how constraining the model by environment specific protein utilization dramatically improved correspondence between predicted and measured growth rates, explaining over 95% of the observed variation.

Yang et al. extended this framework by defining proteome sector constraints that coarse-grain the proteome into functional categories (e.g., translation, transport, energy metabolism) (16). By calibrating ME models with these empirically derived sector allocations, they produced a

"generalist" model that balanced optimal growth with preparedness for stress. Across fifteen growth conditions, the sector-constrained model achieved 69% lower error in growth-rate predictions and 14% lower error in flux predictions relative to the unconstrained model.

Together, these studies established that omics derived constraints provide a formal bridge between experimental data and mechanistic computation. They validated that proteome allocation is not merely an outcome of optimization but an evolved compromise between efficiency and flexibility. Importantly, they also revealed how quantitative omics integration can reduce parameter uncertainty (e.g., k_{eff}values) and guide model refinement.

The same principle extends naturally to transcriptomic summaries such as iModulons. By interpreting omics components as constraints on translation or enzyme usage, models can incorporate real world regulatory context while maintaining mechanistic rigor. This dissertation builds on that foundation: applying iModulon and proteome-derived constraints to explore respiratory plasticity, stress response, and mutation effects, and then encapsulating those methods in a web platform that makes such analyses reproducible and accessible to the broader community.

Chapter 2. Proteome allocation is linked to transcriptional regulation through a modularized transcriptome

It has proved challenging to quantitatively relate the proteome to the transcriptome on a per-gene basis. Recent advances in data analytics have enabled a biologically meaningful modularization of the bacterial transcriptome. We thus investigate whether matched datasets of transcriptomes and proteomes from bacteria under diverse conditions can be modularized in the same way to reveal novel relationships between their compositions. We find that; (1) the modules of the proteome and the transcriptome are comprised of a similar list of gene products, (2) the modules in the proteome often represent combinations of modules from the transcriptome, (3) known transcriptional and post-translational regulation is reflected in differences between two sets of modules, allowing for knowledge-mapping when interpreting module functions, and (4) through statistical modeling, absolute proteome allocation can be inferred from the transcriptome alone. Quantitative and knowledge-based relationships can thus be found at the genome-scale between the proteome and transcriptome in bacteria.

2.1 Background

Omics data types and measurement methods emerged in the late 1990s and early 2000s. Transcriptomes were measured using hybridization to DNA arrays, and proteomes were measured using mass spectrometry. Early attempts to correlate these two omics types were unsuccessful due to complex post-transcriptional and post-translational regulation or to various technical challenges with the measurement technologies (17–19). Later, in the mid to late 2010s, several studies compared the levels of transcripts and protein abundance on a per-gene basis (20–22). Such correlations were achieved for a few transcript-protein pairs in humans and yeast (21,22) but

proved to be more scalable in *Escherichia coli* (20). These studies suggested that correlations between the two omics data types are possible on a small scale.

In the late 2010s, a massive number of RNAseq datasets accumulated for bacterial transcriptomes. This data deluge led to the application of machine learning methods to decompose the bacterial transcriptome into regulatory signals (23). Of these methods, independent component analysis (ICA), a source signal extraction algorithm, was found to modularize the transcriptome into lists of independently modulated genes, termed iModulons (13). A traditional use case of ICA is illustrated in **Figure 2.1a**, where recording devices in a noisy room can discern the sources of noise and their contributions to the measured noise. In a biological context, an expression profile of a given sample is analogous to a microphone, since it is recording combined effects from transcriptional regulators in a noisy environment (**Figure 2.1b**, **Figure 2.1c**). When applied to transcriptomic data, the output of ICA was shown to most successfully match known regulons in a comparison between 42 machine learning methods (23). Moreover, iModulons could be integrated with known binding sites of transcriptional regulators, and compared to their associated regulons (see iModulonDB.org) (24).

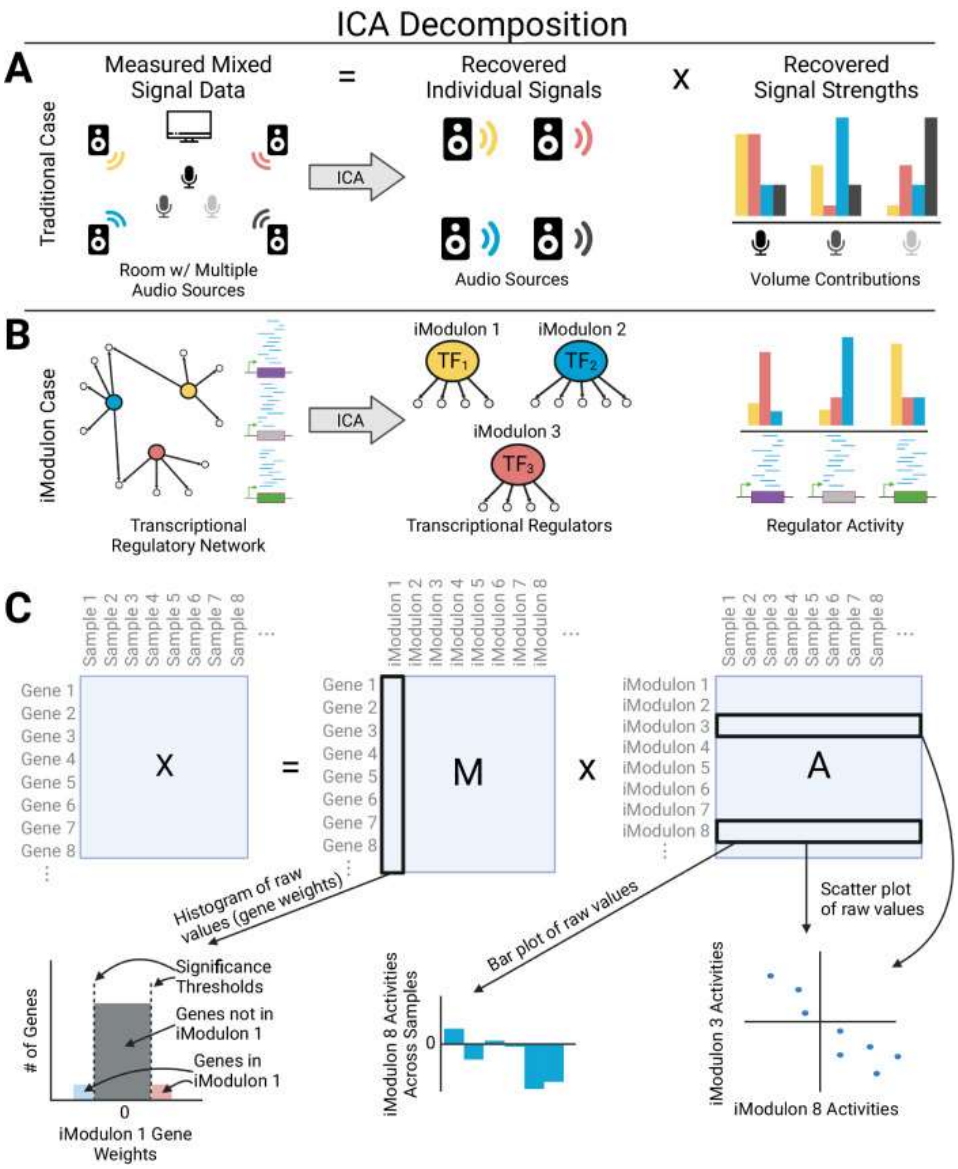


Figure 2.1: Independent component analysis (ICA) extracts individual signals and their strengths from measured mixed signal data.

(A) Three microphones record the sounds in a noisy room with four audio sources. ICA is able to recover the original audio sources and their relative volume contributions to the signal captured by each microphone. **(B)** The transcriptional regulatory network is analogous to the noisy room, and expression data for genes is analogous to the microphones. In this case, ICA is able to recover the transcriptional regulators and their activity that contributes to each gene expression. **(C)** Matrix decomposition representation of the iModulon Case in panel (B). The matrix **X** is your measured mixed signal data (e.g., RNAseq or proteomics). The matrix **M** contains gene weights for all genes. iModulon membership is determined by thresholding the gene weights for all genes in the column. Genes that lie outside the threshold are considered in the iModulon and regulated by the transcription factor noted in panel (B). The matrix **A** contains all the iModulon activity values or the regulator activity level noted in panel (B). A row of **A** can be bar plotted to show the activities across samples, or it can be scatter plotted against another row of **A** to compare iModulon activities.

iModulons from disparate datasets were shown to be similar, indicating that they represent a fundamental decomposition of the transcriptional regulatory network into underlying regulatory signals (25). iModulons could be knowledge-enriched, thus yielding a fundamental understanding of the composition of the transcriptome and how it changes between conditions (26–32). ICA has now been applied to several organisms across the phylogenetic tree (24). This advance led to discoveries of gene functions (33), effects of mutations on protein complex regulation (34), and identifying energetic trade-offs across sample conditions (35). Thus, the knowledge-based modularization of the bacterial transcriptome has led to major advances in understanding its systems characteristics.

This knowledge-based decomposition of the transcriptome naturally leads to the question: can we similarly modularize the proteome? In the present study, we generate and collect proteomic profiles for *E. coli*, modularize this dataset using ICA, and compare the iModulons in the transcriptome to those found in the proteome. This comparison leads to a large-scale, mechanistic interpretation of the relationship between the two omics data types.

2.2 Results

2.2.1 Independent component analysis modularized the proteome

We performed ICA on a compendium of proteomics samples (termed ProteomICA) consisting of 64 proteomes from a previous study (36), and 98 new samples representing conditions matching RNAseq samples in the transcriptomic compendium Precision RNA Expression Compendium for Independent Signal Extraction (PRECISE) (14). These samples contain abundances of 1390 proteins. Since proteomic methods only capture the highest abundance proteins, this is a much lower number than the 4257 genes for which RNAseq finds transcripts (37). The 98 new proteomic samples introduced new growth conditions representing varying stressors,

carbon sources, and supplementations. These new conditions were chosen based on iModulon activities in PRECISE in order to obtain informative matched omics samples that improved signal extraction for ProteomICA (**Figure 2.2a**). ProteomICA has 162 high-quality reproducible proteomes from *E. coli*.

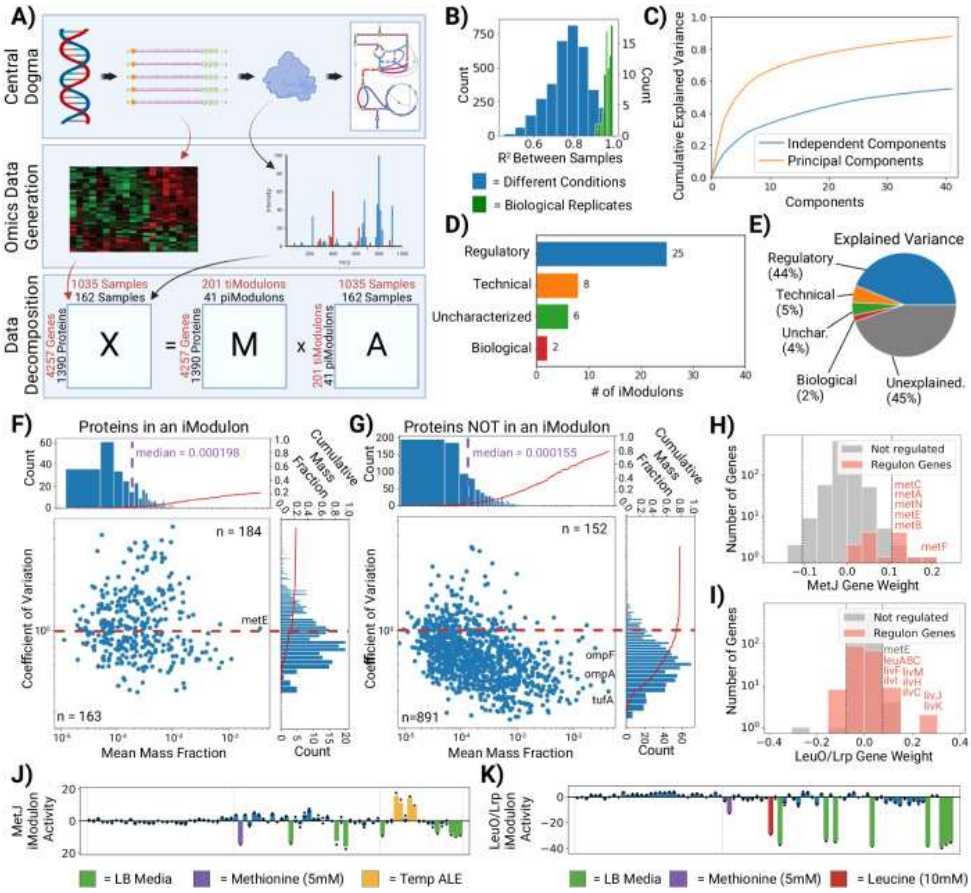


Figure 2.2: ICA of a compendium of proteomic samples (ProteomICA).

(A) The central dogma of molecular biology is the process whereby genetic information is converted into functional proteins that catalyze metabolic reactions and carry out other cellular functions. Genome-wide datasets can be generated for the transcriptome and proteome and analyzed using ICA. Given a matrix of gene expression or of protein abundance, **X**, ICA identifies independently modulated groups of genes or proteins called iModulons (expressed as weights contained in a column of **M**). Every sample in the dataset has an activity associated with each iModulon that becomes condition-specific (row of **A**). Matrix multiplication of **M*A** results in **X**. **(B)** Histogram of Pearson correlations between proteomic samples (biological replicates vs. random samples). **(C)** Cumulative explained variance of the independent components and principal components from matrix decomposition of the proteomics compendium. **(D)** Enrichment categories for proteomic iModulons. **(E)** Pie chart of the explained variances for each enrichment category. **(F), (G)** Scatter plots of the coefficient of variation (CV) and mean mass fractions for proteins in an iModulon, and NOT in an iModulon, respectively. Axes graphs show a histogram of the distribution and the cumulative mass fraction of proteins (red). The dashed red line indicates a CV of 1. Proteins below this threshold are considered invariant. N counts for each section are listed. **(H), (I)** Histogram of the gene weights within the MetJ and LeuO/Lrp independent components (column of **M**), respectively. The significance threshold (gray) identifies the most extreme values, and thus which genes are considered enriched in an iModulon. Regulon genes associated with the iModulon regulator are highlighted. **(J), (K)** proteomic iModulon activity spectrums for MetJ and LeuO/Lrp, respectively. Differentially activated samples are highlighted with sample condition metadata. LB: Lysogeny Broth, Temp: Temperature, Unchar: Uncharacterized.

ProteomICA consists of only high-quality samples with biological replicates having Pearson correlation coefficients greater than 0.90. In contrast, biological replicates in PRECISE have R^2 values greater than 0.95. This difference in reproducibility is in part due to the higher experimental variation in replicate proteome samples as opposed to transcriptome samples (38,39). This difference can also be seen with the higher correlation coefficients between randomly chosen PRECISE samples than between randomly chosen ProteomICA samples (**Figure 2.2b**, **Supplemental Figure A.1**). These characteristics, in turn, with higher technical noise during data generation (40), result in the ProteomICA compendium having a lower overall explained variance from the independent components and principal components than PRECISE (**Figure 2.2c**, **Supplemental Figure A.2**).

The ICA decomposition of the ProteomICA database resulted in 41 proteomic iModulons (piModulons). These piModulons represent the statistically independent protein expression signals found across all 162 samples (81 unique conditions in duplicate) in the ProteomICA compendium. These piModulons represent 25% of detected proteins by count and 22% of the proteome by mass.

The 41 piModulons are classified into different categories (**Figure 2.2d**). We find that 25 of the 41 piModulons correspond to known regulators with well-documented biological functions. Additionally, there are two piModulons that represent a specific biological function without an associated regulator. These two biological piModulons, in conjunction with the 25 regulatory piModulons, explain 46% of the overall explained variance in ProteomICA (**Figure 2.2d**, **Figure 2.2e**). Eight of the remaining 41 piModulons are considered technical and are single gene iModulons, whereas six of the remaining 41 are uncharacterized with no clear function. These final 14 piModulons represent 9% of the overall explained variance in ProteomICA. Thus, taken together, the 41 piModulons explain 55% of the variation in ProteomICA.

Since ICA is a blind source separation algorithm that deconvolutes mixed signals (41), it performs better if the signal strengths vary notably between samples (13,25). We see a higher coefficient of variation (CV) in mass fractions of individual proteins found to be in a piModulon (**Figure 2.2f**) versus those that are not in a piModulon (**Figure 2.2g**). Proteins not in a piModulon account for 78% of the total proteome, with 72% being considered invariant, with CVs less than 1 (n = 891 proteins). In contrast, proteins in a piModulon account for 22% of the proteome with only 47% being considered invariant (n = 163 proteins).

Within the invariant non-piModulon proteins, we see the most abundant protein translation elongation factor, TufA (42), and outer membrane proteins OmpF and OmpA. On the other hand, MetE, coding for homocysteine transmethylase, is a very large protein that catalyzes the final step of methionine biosynthesis in the absence of cobalamin (43), is found in a piModulon due to methionine supplementation conditions that vary its activity. The overall distribution of protein mass fractions is slightly higher for piModulon proteins (median = 0.000198) than proteins not in a piModulon (median = 0.000155).

2.2.2 iModulons are annotated to biologically meaningful functions

The iModulons of the transcriptome have annotated biological functions and most have transcriptional regulators associated with them (iModulonDB.org). The main method for determining the regulatory role of transcriptomic iModulons (tiModulons) is to use the corresponding established regulon in conjunction with the highly weighted genes (in a column of the matrix **M**) to see if there is a significant overlap (13,24). The same approach was used here in the analysis of ProteomICA (**Figure 2.2h**, **Figure 2.2i**). However, due to the small number of samples in ProteomICA compared to PRECISE (162 proteomes vs 1035 transcriptomes, respectively) and fewer proteins than transcripts being identified (1390 proteins vs 4257 genes),

fewer signals are decipherable from the proteomic data. As a result, some piModulons represent a combination of more than one tiModulon.

We illustrate the comparison of the two types of iModulons using two specific examples (**Figure 2.2h**, **Figure 2.2i**). The MetJ piModulon overlaps with the MetJ regulon, and the LeuO/Lrp piModulon overlaps with the Lrp or LeuO regulons. In these two examples, the LeuO/Lrp piModulon consists of the union of the Leucine and Lrp regulons, and the *metE* gene is enriched in both the MetJ and LeuO/Lrp piModulons. The corresponding columns in the iModulon matrix, **M**, contain the weightings for each gene in an iModulon.

The activities for each piModulon are found in the corresponding rows of the matrix **A**. The elements of this row can be used to plot a bar chart that shows the relative activity of the piModulon under a given condition. This bar chart is referred to as the *activity spectrum* for the piModulon (**Figure 2.2j**, **Figure 2.2k**). The activity spectrum for the MetJ piModulon shows that it exhibits low activity in samples with methionine (5 mM) supplementation and LB media, and high activity during low temperatures and adaptive laboratory evolution (ALE) under temperature stress (Temp ALE). The high activities at low temperatures are due to the first step of methionine biosynthesis, homoserine o-succinyltransferases (MetA), being more stable at lower temperatures (44). The LeuO/Lrp piModulon also has low activity in samples with leucine (5 mM) supplementation and LB media, but also with methionine (5 mM) supplementation due to the additional *metE* enrichment. The latter's signal is not as strong due to a lower gene weight of *metE* in the piModulon, and thus, does not see as negative an activity compared to the leucine and LB samples.

Thus, ProteomICA can be decomposed into piModulons using ICA. If there is a corresponding compendium of matched transcriptomic samples available, then the iModulons computed from both can be related to one another (see the following section).

2.2.3. iModulons in the proteome mirror those in the transcriptome

The correlation between iModulon gene weights enables the comparison of the weighted gene content and allows us to match corresponding piModulons and tiModulons computed from PRECISE and ProteomICA (**Figure 2.3a**). We computed Pearson correlation coefficients between gene weights for all pairs of piModulons and tiModulons. pi- and tiModulons are only considered matched if the resulting correlations are above a threshold of 0.25 (**Supplemental Table A.1**). This value is set low due to the non-uniformity of the gene-weight distributions for iModulons with similar functions across organisms or omics data types. Matches were also manually checked. A total of 17 of the 25 regulatory piModulons match with a tiModulon, in addition to both biological piModulons have a matching tiModulon. As mentioned before, some piModulons are combinations of more than one tiModulon. For most of these cases, the piModulon matched with every tiModulon within the combination. For example, the FliA/FlhDC piModulon matched with the FliA and FlhDC-2 tiModulons.

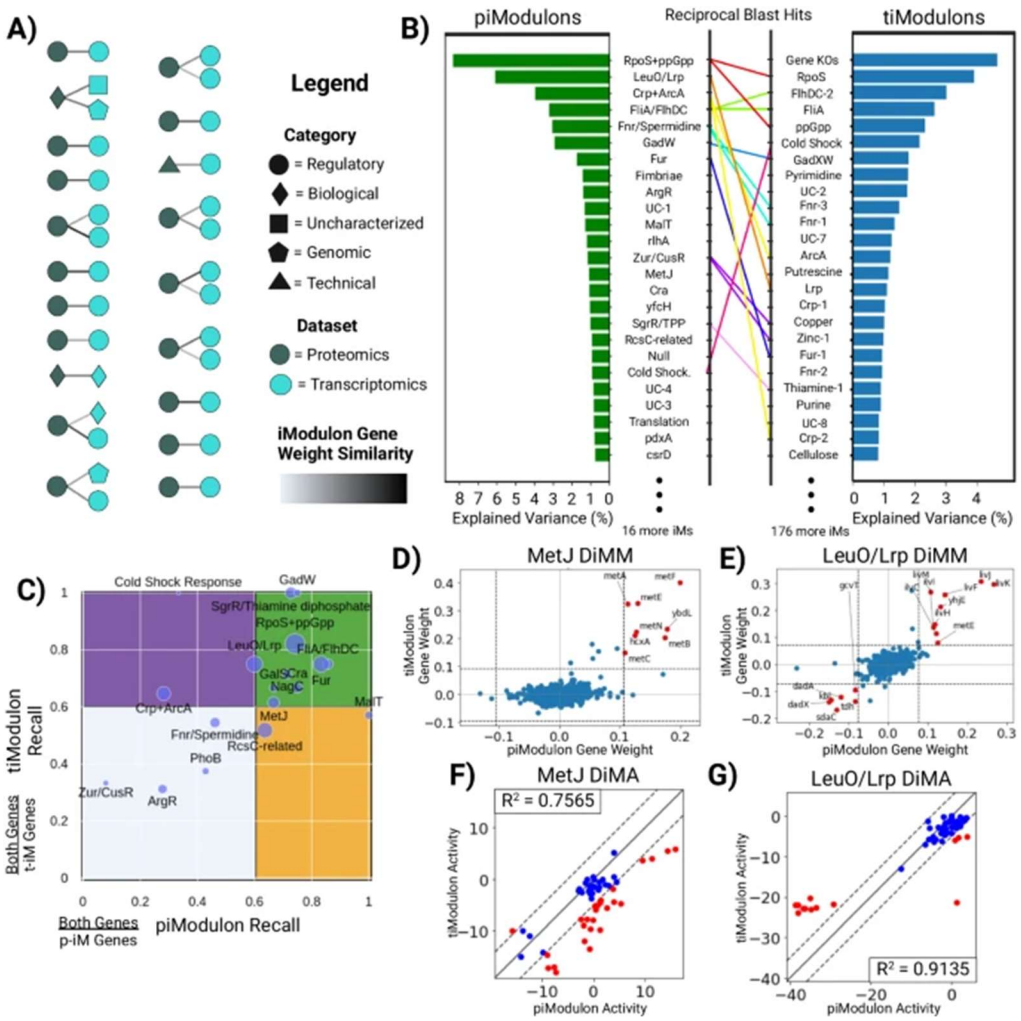


Figure 2.3: Proteomic iModulons (piModulons) exhibit similar gene lists and activity levels as transcriptomic iModulons (tiModulons).

(A) A schematic of correlations between the independent components within the transcriptomic compendium (PRECISE) and the proteomics compendium (ProteomICA). Colors indicate the dataset, while the shapes indicate the enrichment category. The shade of the links is determined by the similarity of the two independent components (Pearson correlation) that match between the two datasets. **(B)** Ranked bar plots of the explained variances for each iModulon within both datasets. Matches between the top-ranked iModulons are shown with rainbow connections. Lumped piModulons that match to multiple tiModulons have multiple connections of the same color. **(C)** Scatter plot of the piModulon recall and tiModulon recall for all matches between the two datasets. 'tiModulon Genes' are genes enriched in the transcriptomic iModulon. 'piModulon Genes' are genes enriched in the proteomic iModulon. 'Both Genes' is the intersection of the tiModulon and piModulon genes. The size of the point is determined by the number of 'Both Genes' **(D)**, **(E)** Differential iModulon Membership (DiMM) plots that compare the gene weights for the matched piModulons and tiModulons for MetJ and LeuO/Lrp, respectively. The significance threshold (gray) shows which genes are enriched in each iModulon. Red genes are enriched in both iModulons. **(F), (G)** Differential iModulon Activity (DiMA) plots that compare the activities for the matched piModulon and tiModulons for MetJ and LeuO/Lrp, respectively. Activities are considered differentially activated for samples that lie outside the significance threshold (gray). Differentially activated samples are highlighted in red.

Upon sorting all iModulons within each compendium by their explained variance and comparing the genes, it becomes evident that there is a very strong correspondence between the iModulon's explained variance between the two omics data types (**Figure 2.3b**). Of the 20 total matches between the datasets, 10 piModulons match to 15 tiModulons that are ranked in the top 25 for both, out of 41 total piModulons and 201 total tiModulons. The 10 piModulons explain 32% of the variability in ProteomICA, while the 15 matched tiModulons explain 26% of the variability in PRECISE. These values are quite similar even with the significant differences in the total number of iModulons obtained from each omic data type. Additionally, the explained variability captured by the two compendia (ProteomICA and PRECISE) is 55% vs 83%, respectively. Thus, the piModulons detect the stronger signals, but cannot detect the more silent signals that the tiModulons can.

Matched iModulon pairs can be described based on the recall they have of each other's gene sets (**Figure 2.3c**). Recall for each matched pair of iModulon can be calculated using the ratio of the number of genes in both iModulons ('Both Genes') to the number of genes in the piModulon or tiModulon (piModulon Recall and tiModulon Recall, respectively). Larger iModulons mostly fall in the high recall green quadrant, whereas smaller iModulons predominantly fall in the light blue low recall quadrant (**Figure 2.3c**). Regulon recall for tiModulons with larger regulons is typically poor (27,28), but that is not observed here, indicating strong correspondence between matched iModulons.

The iModulon matrix **M** and the activity matrix **A** can also be compared for each matched pairs of iModulons. A differential iModulon membership plot (DiMM) compares the gene weights (column of **M**) for matched iModulons between PRECISE and ProteomICA (**Figure 2.3d**, **Figure 2.3e**). Genes enriched in both iModulons are highlighted in red, and ICA is able to identify the same genes in both compendia regardless of gene weight sign. A differential iModulon activity plot

(DiMA) compares the activities for condition-matched samples in PRECISE and ProteomICA (**Figure 2.3f**, **Figure 2.3g**). Correlations between the two activities are calculated with differentially activated samples highlighted in red. Correlations range from strong to weak depending on the number of differentially activated samples and are explored more in the following section.

We thus find that there is good correspondence between matched piModulons and tiModulons, with the former often representing combinations of the latter. The gene composition of matched pi- and tiModulons is congruent, and so are their condition-dependent activities. This correspondence of modularization of the transcriptome and proteome enables deeper analysis.

2.2.4 Matched iModulons reflect established regulatory mechanisms

The ti- and piModulons can be compared in terms of the gene weights (i.e., composition of the signal) and their activity levels (i.e., signal strengths), see **Figure 2.3d-g**. Plotting the differential iModulon activities (DiMA plots) of all pairs of matched ti- and piModulons reveals three distinct groups; pairs that are {1} transcriptome-dominant (signal more active in the tiModulon than the piModulon), {2} proteome-dominant, and {3} neutral. These differences can be interpreted in light of known transcriptional and translational regulation that, in some cases, is condition-specific, but in many cases are broad and well established. All DiMA plots can be found in **Supplemental Figure A.3**. We describe a few cases in detail. The following reviews also provide full descriptions for each type of regulatory mechanism in case readers may not be familiar with them (45–47).

Higher tiModulon activities indicate transcriptional attenuation, riboswitches, or transcript stability that lead to relatively higher RNA than protein: When tiModulon activities are higher than that of the matched piModulon, the iModulon has a stronger signal in the transcriptome and is said

to be transcriptome-dominant. This characteristic can be attributed to transcriptional attenuation, riboswitches that inhibit translation, or stability due to structural reorganization.

LeuO/Lrp (**Figure 2.4a**): The LeuO/Lrp iModulon contains the *leuLABCD* operon, which consists of leucine synthesis genes. The operon is known to be regulated by ribosome-mediated attenuation in the presence of charged leucine tRNAs (48). Thus, we observe this mechanism as a transcriptome-dominant iModulon: in rich media or leucine-supplemented minimal media, the expressed RNA is not translated, leading to an upregulation of the tiModulon relative to the piModulon. We also observed the iModulon becoming proteome-dominant in the case of arginine supplementation, which may indicate competitive repression by arginine of the dual regulator Lrp and its associated operons.

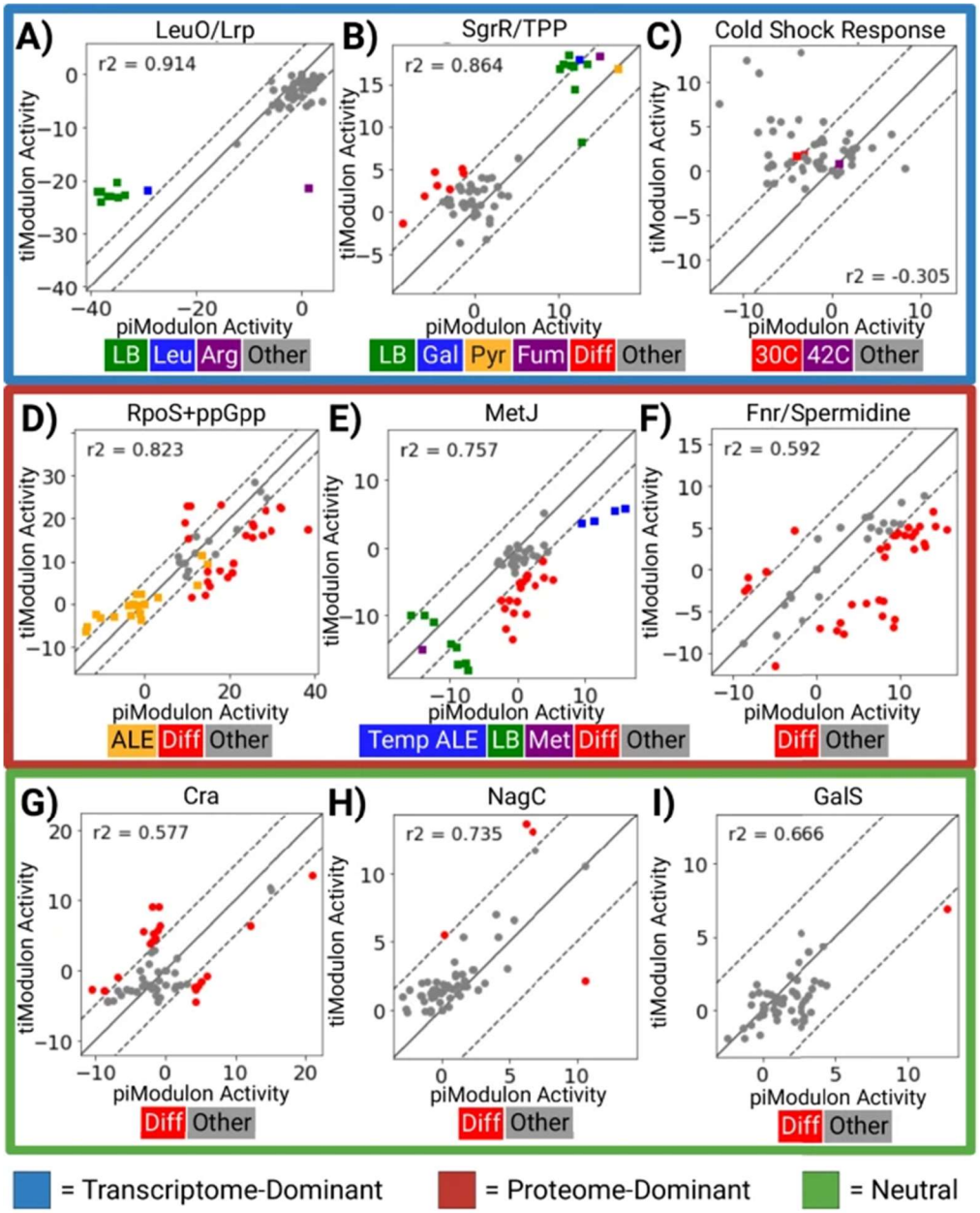


Figure 2.4: Comparing piModulon and tiModulon activities for matched samples reveal condition-specific regulatory mechanisms.

Differential iModulon Activity (DiMA) plots for some matched iModulons between the two datasets. Plots are categorized by the observed result due to regulatory mechanisms, such as riboswitches, transcriptional attenuation, temperature-dependent transcript structural reorganization, and protein product autoregulation. **(A–C)** iModulons that are transcriptome-dominant (signal more active in the tiModulon than the piModulon) are highlighted in blue. **(D–F)** iModulons that are proteome-dominant are highlighted in red. **(G–I)** iModulons that are neutral are highlighted in green. Activities are considered differentially activated for samples that lie outside the significance threshold (dashed line). Legends for each plot are placed below each plot. LB: Lysogeny Broth, Leu: Leucine Supplement, Arg: Arginine Supplement, Gal: Galactose Carbon Source, Pyr: Pyruvate Carbon Source, Fum: Fumerate Carbon Source, Diff: Differentially Activated, Temp: Temperature, Met: Methionine Supplement. $n = 57$ conditions.

SgrR/Thiamine diphosphate (**Figure 2.4b**): Thiamine diphosphate (TPP) can act as a ligand that binds to a riboswitch that inhibits the translation of the *thiMD* and *thiCEFSGH* operons (49,50). Most samples in rich LB media are differentially active towards their respective tiModulon activities, probably due to thiamine-induced premature transcriptional termination. Additionally, samples grown on galactose, pyruvate, and fumarate have higher overall pi- and tiModulon activities due to increased demand for the thiamine cofactor in essential reactions pyruvate dehydrogenase complex and 2-oxoglutarate (2-ketoglutarate) dehydrogenase complex.

Cold Shock Response (**Figure 2.4c**): At temperatures below 37 °C, the *cspA* mRNA undergoes temperature-dependent structural reorganization (51). This structural change is likely due to the stabilization of an otherwise thermodynamically unstable folding intermediate. At low temperatures, the structure is also less susceptible to degradation (52). Samples at 30 °C are differentially active and transcriptome-dominant, while samples at 42 °C have no activity due to mRNA instability at high temperatures.

Higher piModulon activities indicate translation activation, protein product autoregulation, or riboswitches that lead to relatively higher protein than RNA: When piModulon activities are higher than that of the matched tiModulon, the iModulon has a stronger signal in the proteome and is said to be proteome-dominant. This characteristic can be attributed to riboswitches that promote translation, protein products autoregulating transcription, or transcript inhibition due to other proteins.

RpoS+ppGpp (**Figure 2.4d**): RpoS, the major stress-related sigma factor, and guanosine 3,5-bispyrophosphate (ppGpp), an important alarmone, both act as master regulators for a wide range of genes including those involved in oxidative stress, temperature shock, acid stress, starvation, and osmotic stress (53). Both regulators integrate several stress signals, and ppGpp helps stabilize RpoS, leading to complex transcriptional regulation (53). In addition, ppGpp was recently

found to directly activate the translation of some genes (54). Thus, we observe a proteome-dominant expression pattern for this iModulon in several samples. Samples from ALE are also known to have low RpoS activities, which is replicated here with both pi- and tiModulon activities (13).

MetJ (**Figure 2.4e**): MetJ regulates methionine synthesis genes at the transcriptional level in response to methionine and related molecules (55). As expected, both the tiModulon and the piModulon are therefore downregulated in LB media and with methionine supplementation. One member of this iModulon, MetA (homoserine o-succinyltransferase, the first step of methionine biosynthesis), is inherently unstable under stressful conditions and high temperatures (44). Thus, it is regulated by temperature-dependent proteolysis (56). Interestingly, conditions which make the protein more stable, such as low temperatures and heat-tolerant strains with *metA* mutations (Temp ALE), are proteome-dominant for this iModulon. This observation likely reflects the increased stability of the MetJ-regulated proteins in those conditions.

Fnr/Spermidine (**Figure 2.4f**): Spermidine is a known small molecule that binds to a riboswitch that facilitates the translation of the *oppABCDF* operon (57). Like the other polyamines, putrescine, and spermine, it stimulates the assembly of 30S ribosomal subunits, increasing general protein synthesis up to 2-fold (58). Here, we see more samples with higher piModulon activities than tiModulon activities due to this increase.

ti- and piModulons that contain their regulator show similar signal strengths in the transcriptome or proteome: When tiModulon activities are similar to that of the matched piModulon, the ti- and pi-iModulons have similar strengths in the transcriptome and proteome and is said to be neutral. This can be illustrated by looking at the Cra, NagC, GalS iModulons (**Figure 2.4g-i**). The DNA-binding transcriptional dual regulator Cra is a member of both the Cra ti- and piModulon. The GlcNaC tiModulon contains its transcriptional regulator NagC, as well as the

Galactose tiModulon which contains GalS. It's an uncommon phenomena for an iModulon to contain its own regulator because regulators typically don't exhibit linear relationships with the genes they regulate due to the complexity of the TRN. If an iModulon exhibits unique characteristics, such as temporal dynamics, the regulatory function is split into multiple iModulons in which case one will contain the regulator and the others won't (e.g., Phosphate-1,2; FlhDC-1,2,3; NtrC-1,2,3; Fnr-1,2,3) (14,24). When the regulator is in an iModulon of a single function that didn't split, it indicates that there aren't any complex regulatory interactions since the iModulon activity is correlated with its regulator's expression. In the case of Cra, NagC, and GalS, this leads to an overall neutral relationship between the proteome and transcriptome as seen here.

Previous studies have clearly shown that tiModulons can be knowledge-enriched by mapping known transcription factor binding sites in promoters of genes found in a tiModulon (24). The results presented in this section take knowledge enrichment a step further. Namely, various molecular mechanisms are reflected in the relative activity levels of pi- and tiModulons. Thus, the ability of ICA to detect these regulatory mechanisms enables us to knowledge-enrich the relationships between iModulons. When piModulon and tiModulon activities are not well-correlated, they indicate post-transcriptional regulatory events. Many such events have been previously characterized in the literature, as described in this section.

2.2.5 tiModulon activities allow prediction of proteome allocation

Revealing the relationships between tiModulons and piModulons opens up the possibility of predicting the composition of the proteome straight from RNAseq data. Such predictions would be advantageous since the composition of the transcriptome can be measured cheaper, faster, and with higher precision and accuracy than the composition of the proteome.

We thus sought to find quantitative relationships between RNAseq data and proteome allocation. Three types of relationships (linear, exponential, broken line, **Figure 2.5a**) were

identified by plotting tiModulon activities against the mass fraction of the proteome allocated to the genes represented by the tiModulon. tiModulon activities that are linearly correlated with their proteome allocation indicate proteome-optimized sectors (i.e., amino acid biosynthesis). tiModulon activities that are exponentially correlated with their proteome allocation indicate proteome-optimized, yet expensive sectors (i.e., stress-related responses). Finally, tiModulon activities that fit a broken line indicate a thresholding response due to phenomena like bet-hedging (e.g., central carbon metabolism) (15,59–63).

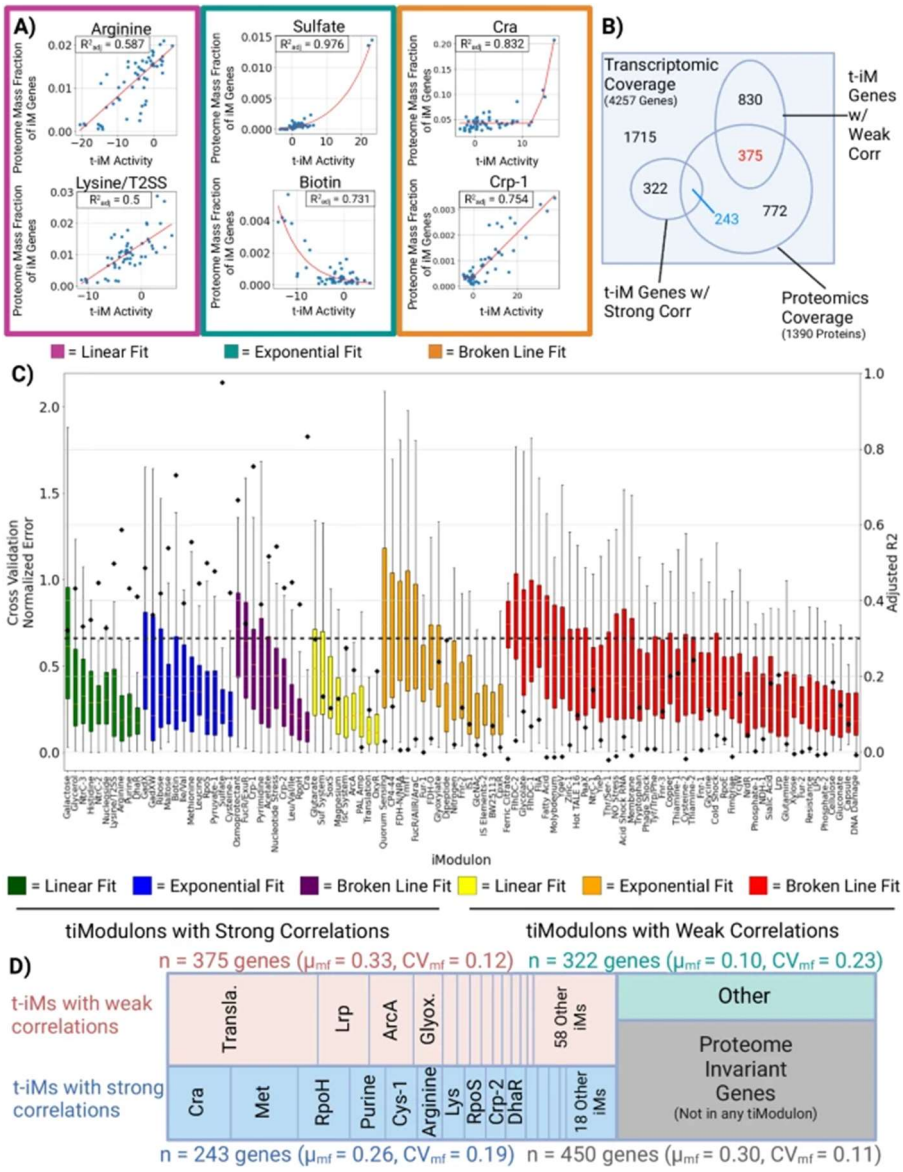


Figure 2.5: Predictability of proteome allocation using tiModulon activities.

(A) Scatter plots for selected tiModulon activities and their measured proteome mass fraction of the associated enriched genes. tiModulons are characterized based on which regression method resulted in the best adjusted R^2 value. The three fits were linear, exponential, and broken line. **(B)** Venn diagram detailing the number of genes/proteins covered by both datasets, in addition to the regression results. A tiModulon is considered to have a strong correlation with its proteome allocated if the adjusted R^2 value ≥ 0.3. tiModulon genes covered by proteomics data with strong correlations are in blue, while covered tiModulon genes with weak correlations are in red. **(C)** Boxplots showing the distribution of normalized errors after cross-validation ($\frac{|ypred-ytest|}{yavg}$). Final model adjusted R^2 values are scattered on top of the boxplots with black diamonds. iModulons are organized/colored by their fitting method and quality of the fitting. $n = 57$ conditions for each model. **(D)** Treemap of the proteome allocation using the tiModulon regression results. tiModulons with strong correlations are in blue, while those with weak correlations are in red. Genes that are not in a tiModulon and whose protein has a CV ≤ 1, are invariant and labeled in gray. Genes that are not in any tiModulon nor proteome invariant are labeled in green.

tiModulon activities that represent strong correlations with their proteome allocation account for 565 genes, 243 of which are covered by ProteomICA (**Figure 2.5b**). tiModulon activities are considered to have a strong correlation with proteome allocation if the adjusted R^2 of the regression fitting after leave-one-out cross-validation is above 0.3 (**Figure 2.5c**). The adjusted R^2 0.3 cutoff was selected as a corrected threshold to use between the three fitting types, as it is equivalent to an R^2 of 0.7 for linear regression in this dataset. These 243 genes account for, on average, 26% of the proteome in the compendia, with a CV of 0.19 (**Figure 2.5d**). All of the scatter plots and regressions for tiModulons with strong correlations can be found in **Supplemental Figure A.4**. tiModulon activities that represent weak correlations with their proteome allocation account for 1205 genes, 375 of which are covered by ProteomICA (**Figure 2.5b**). These genes account for, on average, 33% of the proteome in the compendia, with a CV of 0.12 (**Figure 2.5d**). Proteome invariant genes that are not in a tiModulon account for 30% of the proteome (CV = 0.11), and genes that are not in a tiModulon but are not invariant account for 10% of the proteome (CV = 0.23).

We also wanted to ensure that these results and methods were robust, so we did additional analysis on our regression methods with holdout splitting percentages ranging from 10% to 30% in 5% increments on top of leave-one-out cross-validation (**Supplemental Figure A.5**, **Supplemental Figure A.6**).

Being able to infer absolute proteome allocation from the transcriptome alone, regardless of condition, requires generalizable statistical models with large adjusted R^2 values. Normalized cross-validation error for each regression is not statistically significant between strongly and weakly correlated tiModulons, yet the adjusted R^2 values differ quite substantially due to outliers in both datasets that cause large errors. While these weaker regressions cannot be used for generalization, unlike the stronger regressions, some can still be used to estimate the proteome allocated for the

specific conditions that do not fall in outlier conditions. For example, the translation tiModulon has one of the weakest correlations but accounts for 11% of the proteome, but due to a small number of outliers is categorized as weak (**Supplemental Figure A.7**). Removal of the outlier conditions would categorize the tiModulon as strong and enable inference of proteome allocation.

Taken together, these results show that ICA decomposition of the transcriptome enables inference for 56% of the proteome allocation for general cases (gray and blue sectors, **Figure 2.5d**), with up to an additional 33% being inferrable for specific conditions (red sector, **Figure 2.5d**). These relationships include the effects of post-transcriptional regulation and should represent practical ways of estimating how differential regulation of gene expression affects the proteome composition. These results provide a strong impetus for generating larger proteomic datasets to generate stronger and broader correlations between the datasets and proteome allocation.

2.3 Discussion

Recent advances in big data analytics have enabled the knowledge-enriched modularization of the transcriptome for various microbial species (13,26–32). Here we investigated if matched datasets of transcriptomes and proteomes could be modularized in the same way to reveal novel relationships between their compositions. Using ICA analysis of matched datasets we found that; (1) the modules of the proteome and the transcriptome are comprised of a similar list of gene products, (2) the modules in the proteome often represent combinations of modules from the transcriptome, (3) known transcriptional and post-translational regulation is reflected in differences between two sets of modules, allowing for knowledge-mapping when interpreting module functions, and (4) through statistical modeling, absolute proteome allocation can be inferred from the transcriptome alone.

Modularizing the proteome via ICA decomposition has resulted in biologically meaningful groups of independently modulated genes, termed proteomic iModulons, or piModulons. They are similar to previous studies that have successfully modularized the transcriptome for various organisms using ICA (24). We show that the piModulons have a similar gene composition as transcriptomic iModulons or tiModulons. While the proteomics compendium used, ProteomICA is newer and has five times fewer samples than the transcriptomics compendium used, PRECISE (14), the former produces detectable signals in just under five times the number of independent modules computed from the latter. This result suggests that if we expand and improve the quality of ProteomICA, perhaps with the inclusion of additional post-translational modifications in search parameters, we may be able to achieve a higher fidelity understanding of the regulation of proteome allocation.

Due to this size limitation, a number of identified piModulons from ProteomICA represent combinations of tiModulons. While this may seem problematic at first, a similar phenomenon is visible when decomposing the transcriptome at lower dimensionalities (64), and it has been shown that iModulons tend to split as more conditions are added, enabling ICA to identify more signals in the datasets (13). It is quite promising to see that a number of these combined modules in the proteome represent the highest explained variance in the compendia of both data types. PRECISE has a total of 201 tiModulons that explain 83% of the variance in the dataset, while ProteomICA has a total of 41 piModulons that explain 55% of the variance in the dataset, but the top-ranked iModulons that are matched between both represent 26% and 32% of the variance, respectively. Note that ICA-derived explanations are based on knowledge, or mechanisms, in contrast to the explanation of statistical variation that is obtained using principal component analysis (PCA).

The congruence of the gene compositions of matched ti- and piModulons led to the comparisons of their activity levels. Such comparison enabled further knowledge enrichment of the

matched sets of iModulons over and above their individual annotation with regulatory knowledge. The comparison allowed the attribution of a number of established transcriptional and post-translational regulatory mechanisms. Regulatory phenomena, such as riboswitches and attenuation, are easily identifiable when comparing matched iModulons of the corresponding regulatory component. Thus, interoperable data analytics at the genome-scale can capture an increasing number of established regulatory mechanisms through detailed molecular biology studies.

We also showed that it is possible to utilize transcriptomic datasets to infer proteome allocation. Previously, this was only achievable on a per-gene basis, but modularization via ICA has scaled up the scope to sets of genes enabling inference of proteome re-allocation (20–22). Transcriptomic samples can be measured cheaper, faster, and with higher precision and accuracy than proteomics samples. The fact that we can demonstrate a correlation between tiModulon activity levels and proteome allocation with this method in its infancy, provides a strong impetus to explore how broadly we can achieve this correlation which requires the generation of larger matched sets of transcriptomic and proteomic datasets. As we generate more matched sets, our regression models will become increasingly more robust since we only currently have 57 matched conditions for regression. In its current form, we train the regression models on all data points using leave-one-out cross-validation due to our dataset size limitation. However, the inclusion of additional holdouts (**Supplemental Figure. A.5**, **Supplemental Figure 6**) doesn't significantly deteriorate the proteome allocation prediction capacity and in fact, further supports the robustness of our initial models. More data will enable more rigorous validation testing of these regression models in follow-up studies.

Furthermore, enabling the prediction of proteome re-allocation between conditions using transcriptomics can further bridge the gap between observable physiological states and molecular profiling methods. Genome-scale computational models that compute proteome allocation can now

be parameterized better and thus be used to build quantitative relationships between the regulation of gene expression and physiological functions and fitness (6,16).

Taken together, we have shown that ICA can modularize the transcriptome and proteome in a consistent manner. The iModulons are knowledge-enriched and thus interpretable based on the fundamentals of cell and molecular biology. This achievement enables the meaningful interoperability of two key omics data types, leading to quantitative and knowledge-based relationships at the genome-scale between the proteome and transcriptome. This capability, in turn, gives us a deep understanding of the systems biology of bacteria, which leads to interpreting their adaptation and changes to environmental stimuli. Thus, distal and proximal causation can be studied at a new scale to more deeply understand organism fitness and survival strategies.

2.4 Methods

2.4.1 Proteomic sample preparation

Frozen cell pellets were resuspended in lysis buffer (75 mM NaCl (Sigma–Aldrich), 3% sodium dodecyl sulfate (Fisher Scientific), 1 mM sodium fluoride (VWR International, LLC), 1 mM β-glycerophosphate (Sigma–Aldrich), 1 mM sodium orthovanadate, 10 mM sodium pyrophosphate (VWR International, LLC), 1 mM phenylmethylsulphonyl fluoride (Fisher Scientific), 50 mM HEPES (Fisher Scientific) pH 8.5, and 1× complete EDTA-free protease inhibitor mixture). Samples were vortexed and sonicated (Qsonica, Q500 equipped with a 1.6-mm microtip) at 20% amplitude for three cycles of 2 s of sonication followed by 2 s of rest, with a total sonication time of 12 s.

Total protein abundance was determined using a bicinchoninic acid Protein Assay Kit (Pierce) as recommended by the manufacturer. Six micrograms of protein were aliquoted for each sample. Sample volume was brought up to 20 µL in a solution of 4 M Urea and 50 mM HEPES,

pH = 8.5. Proteins were reduced and alkylated with 5 mM dithiothreitol (DTT) for 30 minutes at 56 °C and 15 mM iodoacetamide (IAA) at room temperature in the dark for 20 min. The reaction was quenched with the addition of 5 mM DTT for 15 min at room temperature in the dark. Proteins were precipitated by adding 5 uL of 100% trichloroacetic acid on ice for 10 min, then centrifuged at 16,000 × *g* for 5 min at 4 °C. The supernatant was removed, and pellets were washed gently in 50 uL of ice-cold acetone. The wash was repeated twice, and the pellets were dried on a heating block at 56 °C. Pellets were resuspended in 1 M Urea and 50 mM HEPES, pH 8.5. The UPS2 Standard (Sigma) was reconstituted as follows: 20 µL of 4 M Urea and 50 mM HEPES, pH 8.5 was added to the stock tube and vortexed and sonicated for 5 min each. Reduction and alkylation were performed as described above. The standard was then diluted in 50 mM HEPES, pH 8.5 such that the final concentration of urea was 1 M. Then 0.88 µg of the standard was spiked into each experimental sample. Samples were then digested first with 6.6 µg of LysC at room temperature overnight followed with 1.65 ug sequencing grade trypsin (Promega) for 6 hours at 47 °C. Digestion was terminated with the addition of 3.3 µL 10% trifluoroacetic acid (TFA) and was brought to a final volume of 300 uL with 0.1% TFA. Samples were centrifuged at 16,000 × *g* for 5 min and desalted with in-house-packed Stage-Tips (36,65). Samples were then dried in a speedvac, and stored at −80 °C until LC–MS/MS.

2.4.2 LC-MS/MS

Samples were resuspended to 1 µg/µL in 5% acetonitrile (ACN) and 5% formic acid (FA), vortexed, and sonicated. Samples were analyzed on an Orbitrap Fusion Mass Spectrometer with in-line Easy NanoLC (Thermo) in technical triplicate. Samples were run on an increasing gradient from 6 to 25% ACN + 0.125% FA for 75 min, then 100% ACN + 0.125% FA for 10 min. One microgram of each sample was loaded onto a 35 cm length in-house−pulled and −packed glass capillary column (ID 100 µm, OD 360 µm) heated to 60 °C. The column was triple packed first

with C4 resin (5 µm, 0.5 cm, Sepax), then C18 resin (3 µm, 0.5 cm, Sepax), and finally C18 resin (1.8 µm, 29 cm, Sepax). Electrospray ionization was achieved through application of 2000 V to a stainless-steel T-junction connecting the sample, waste, and column. The mass spectrometer was run in positive polarity mode with MS1 scans performed in the orbitrap (375 m/z to 1500 m/z, 120,000 resolution, AGC set to 5×10^5, ion injection time of 100 ms maximum, dynamic exclusion set to 30 s duration). Top N was used for fragment ion isolation, with N set to 10. A decision tree was used to isolate ions with a charge state of two between 375 m/z and 1500 m/z, and ions with charge states of 3–6 were isolated between 600 m/z and 1500 m/z. Precursor ions were fragmented using fixed collision-induced dissociation and fragment ions were detected in the linear ion trap in profile mode. Target AGC was set to 1×10^4.

Technical triplicate spectral data was searched against custom reference proteomes of the respective strains (see above) with the UPS2 database appended using Proteome Discoverer 2.5 (Thermo). Spectral matching and an in-silico decoy database were performed using the SEQUEST algorithm (66). Precursor ion mass tolerance was set to 50 PPM, and fragment ion tolerance was set to 0.6 Daltons. Trypsin and LysC were specified as digesting enzymes with a maximum missed cleavage of two sites allowed. Peptide length was set between 6 and 144 amino acids. Dynamic modifications included the oxidation of methionine (+15.995 Da), and static modifications included carbamidomethylation of cysteines (+57.021 Da). A false discovery rate of 1% was applied during spectral searches.

2.4.3 Proteome abundance estimations

The protein abundance estimation steps used on the new dataset are the same used on the previous PXD015344 (36). The top3 metrics were calculated for each protein as the average of the three highest peptide areas (21,67). Linear regression was used to calibrate the top3 metric with the UPS2 standard according to the following model:

$$\log_{10}(A) = a + b\log_{10}(top3)$$

Where A is the amount of loaded protein A and top3 is the average of the three highest peptide areas. In order to obtain abundance relative to cell dry weight, we used the following formula:

$$C_i = \gamma \frac{A_i}{\sum_j A_j}$$

Where the numerator of the ratio, Ai, is the abundance of the ith protein, and the denominator is the sum of abundances for all j proteins. We use a constant ratio $\gamma = 13.94$ umol*gDW^{-1} (68).

2.4.4 Compiling ProteomICA and data imputation

Upon estimating protein abundances, proteins with <50% coverage within the dataset were removed. Of the remaining proteins, samples with no abundances were replaced with the minimum global protein abundance. Datasets were then converted to mass fractions or protein concentrations and concatenated to compile the proteomics compendia. Similar to how the final transcriptomics expression compendium is log-transformed $\log_2(TPM + 1)$, the final proteomics expression compendium is scaled by a million and also log-transformed similarly $\log_2(PPM + 1)$ (13). Biological replicates with $R^2 < 0.9$ were removed to reduce technical noise. Individual datasets were then centered using a common reference condition between all datasets to reduce batch effects.

2.4.5 Independent component analysis

ICA was run following the PyModulon workflow. ICA is implemented using the optICA extension of the popular algorithm FastICA. The script can be found (https://github.com/SBRG/iModulonMiner/tree/main/4_optICA). The output of the algorithm are

two matrices, **M** and **A**, given an input matrix **X**. In our case, the matrix **X** is our curated proteomics compendium. The matrix **M** contains robust independent components, and the matrix **A** contains their corresponding activities.

2.4.6 Computing independent components and their enrichments

After running ICA and obtaining the resulting matrix decomposition, the PRECISE1K workflow (https://github.com/SBRG/precise1k) (14) is used to choose the optimal dimensionality of the resulting ICA runs and associate regulator enrichments to iModulons. After automation, enriched iModulons are checked for their associated regulators and uncharacterized iModulons are manually curated using a variety of annotation tools such as COG and GO terms.

2.4.7 Comparing iModulons between PRECISE and ProteomICA

The PyModulon python package (https://github.com/SBRG/pymodulon) (13) was used for the DiMM, and DiMA, and explained variance plotting functions. The PyModulon package also enables comparison between organisms via the compare_ica function, which was utilized to correlate the iModulons between PRECISE and ProteomICA. The compare_ica function uses only the overlap of the two gene sets to calculate correlations, and has a default threshold of 0.25 which was not changed. The ICA workflow was run on a subset of PRECISE that contained only the 1390 genes covered by proteomics and 137 matched samples (57 conditions without replicates), centered using the same common reference condition as ProteomICA. Population samples were not included in the matched dataset, only clones. The resulting decomposition was used for DiMM, DiMA, and recall plots.

2.4.8 Regression and cross-validation for proteome allocation

Only matched samples that are present in both PRECISE and ProteomICA were used for this analysis. Uncharacterized, Genomic, and Technical tiModulons were ignored. For each tiModulon, tiModulon activity was plotted against its associated proteome mass fraction. Replicates were averaged for both iModulon activity and proteome allocation. Leave-one-out cross-validation was performed with three different fits, linear, exponential, and broken line. The model with the lowest mean average error was selected as the best fitting method for that tiModulon. To analyze the robustness of the models and results, data was split into train and test groups, with the test group ranging from 10% to 30% in increments of 5%. Leave-one-out cross-validation was conducted on the training set, with the final model parameters evaluated against the test set.

2.4.9 Calculating proteome allocated to groups of tiModulons

Proteome allocation to groups of tiModulons was calculated sequentially to avoid multiple iModulon gene memberships. First, tiModulons with strong correlations were sorted in descending order by model performance. tiModulon gene lists were extracted and used to calculate the mean and CV for mass fractions across the compendia. After which, the genes were removed from the total gene list (1390 genes in the proteome coverage) and could no longer be used for another tiModulon. tiModulons with weak correlations were calculated in a similar fashion after iterating through all the tiModulons with strong correlations. Lastly, proteome invariant genes and then 'Other' were calculated straight from the remaining gene list since there could no longer be overlap.

2.4.10 Statistics and reproducibility

No statistical method was used to predetermine size. Two samples were excluded from analysis due to not meeting the biological replicated threshold (see Compiling ProteomICA and

data imputation section). The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment.

2.4.11 Data and code availability

All MS-based proteomics raw files for newly run samples are available on the ProteomeXchange Consortium with the dataset identifier PXD039558. All previously used proteomics data can be found under the identifier PXD015344. All data generated or analyzed during this study are included in the published article (and its Supplementary Information Files). The full PRECISE1k decomposition matrices are available at https://github.com/SBRG/precise1k. Raw RNA-seq data used in PRECISE1k have been deposited at GEO and are publicly available. Accession numbers are listed in the metadata file located in the GitHub repository at the path: data/precise1k/metadata_qc.csv.

Code for our Independent Component Analysis pipeline can be found on GitHub (https://github.com/SBRG/iModulonMiner, https://github.com/SBRG/precise1k). Code for all iModulon analysis can also be found on GitHub (https://github.com/SBRG/pymodulon).

2.5 Acknowledgements

The work was funded by the Novo Nordisk Foundation Grant Number NNF20CC0035580, the National Institute of General Medical Sciences of the National Institutes of Health Grant R01 GM057089, and by the generous support of the Y.C. Fung Endowed Chair. We would like to thank Daniel Zielinski and Cameron Lamoureux for their helpful discussions. We would also like to thank Marc Abrams for their help with manuscript proofreading. Figures were created using Biorender.com.

A.P., A.V.S., and B.O.P. designed the study. D.M., Y.H., A.C., D.J.G., and S.M. performed experiments. A.P. analyzed the data. A.P., K.R., and B.O.P. wrote the manuscript with contributions from all other co-authors.

Chapter 2, in part, is a reprint of the material as it appears in *Nature Communications*: **Patel A**, McGrosso D, Hefner Y, Campeau A, Sastry AV, Maurya S, et al. Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. *Nat Commun.* 2024 Jun 19;15(1):5234.

The dissertation author was the primary author.

Chapter 3. Respiratory plasticity and the aerobicity stimulon in *E. coli* revealed by laboratory evolution and multi-scale computational systems biology of electron transport system variants.

Respiration in *Escherichia coli* provides a model system for examining how metabolic networks reorganize to balance energy efficiency and proteome allocation under diverse conditions. In this chapter, we investigate respiratory plasticity using strains engineered with unbranched electron transport systems (ETS) that differ in proton-pumping efficiency. Two complementary manuscripts are presented. The first establishes how distinct ETS configurations adapt through laboratory evolution, revealing that alternative pathway rewiring enables all variants to converge on optimized growth and defining the Aero-Type System (ATS) as a generalized framework for aerobic bioenergetics. The second extends this approach to nutrient-limited conditions, integrating multi-omics analysis and genome-scale modeling to identify a conserved aerobicity stimulon and carbon source-specific adaptive mechanisms. Together, these studies exemplify the dissertation's overarching theme by showing how knowledge-enriched data analytics and modeling tools interoperate to generate a systems-level understanding of respiratory plasticity and energy homeostasis.

3.1 Background

Respiration requires organisms to have an electron transport system (ETS) for the generation of proton-motive force across the membrane that drives ATP synthase. Although the molecular details of the ETS are well-studied and constitute textbook material, few studies have appeared to elucidate its systems biology. The most thermodynamically efficient ETS consists of two enzymes, an NADH: quinone oxidoreductase (NqRED) and a dioxygen reductase (O_2RED), which facilitate the shuttling of electrons from NADH to oxygen. However, evolution has produced variations within the ETS which modulate the overall energy efficiency of the system even within

the same organism (69–71). The systems level impact of these variations and their individual physiological optimality remain poorly determined. To mimic varying ETS efficiency we generated four *Escherichia coli* deletion strains (named ETS-1H, 2H, 3H, and 4H), each with one of the four unbranched ETS variants that pump 1, 2, 3, or 4 proton(s) per electron, respectively. We then performed systems level characterization of these ETS variants. We observe that: (a) adaptive laboratory evolution (ALE) enables all four ETS variants to evolve to a similar growth rate; (b) the evolution of ETS variants is supported by specific rewiring of major energy-generating pathways that couple to the ETS to optimize their ATP production capability; (c) proteome allocation per ATP generated is the same for all the variants, (d) the aero-type, that designates the overall ATP generation strategy (72) of a variant, remain conserved during its laboratory evolution, with the exception of the ETS-4H variant; and (e) integrated computational analysis of the data supports a proton-to-ATP ratio of 10 protons per 3 ATP for ATP synthase for all four ETS variants.

3.2 Results and Discussion

E. coli has a highly flexible ETS consisting of 15 dehydrogenases and 10 reductases to allow growth in both oxic and anoxic environments (73). The expression of these enzymes is regulated by a variety of electron acceptors with a known hierarchy, such that oxygen represses all anoxic respiratory pathways and nitrate represses other anoxic pathways (71,73). Despite this thermodynamic hierarchy, co-expression of different respiratory chains was reported in another γ-proteobacterium to expand the flexibility of its electron transfer network (71). We probed the condition-dependent expression of all these dehydrogenases and reductases using a large RNA-seq compendium for *E. coli* (14). We observed a spectrum of expression values of these genes across the experimental conditions showing the contribution of these enzymes in generating plasticity in energy metabolism (**Supplemental Figure B.1**).

To examine the contributions of individual oxic respiratory pathways to bioenergetics, we sought to design unbranched pathways through the ETS. The oxic component is contributed by both proton pumping and non-pumping NqREDs (hereafter referred to as NDH-I and NDH-II, respectively) along with three types of O_2REDs (**Figure 3.1a**). Cytochrome bd O_2REDs (CBDs) are less electrogenic compared to Cytochrome bo_3 O_2REDs (CYO). There are two CBDs, bd-I and bd-II, and both functions similarly to generate proton-motive force (PMF) by a vectorial movement of protons involving transmembrane charge separation. The similar PMF generation strategies make bd-I and bd-II O_2REDs equivalent when choosing gene knockout strategies (74,75).

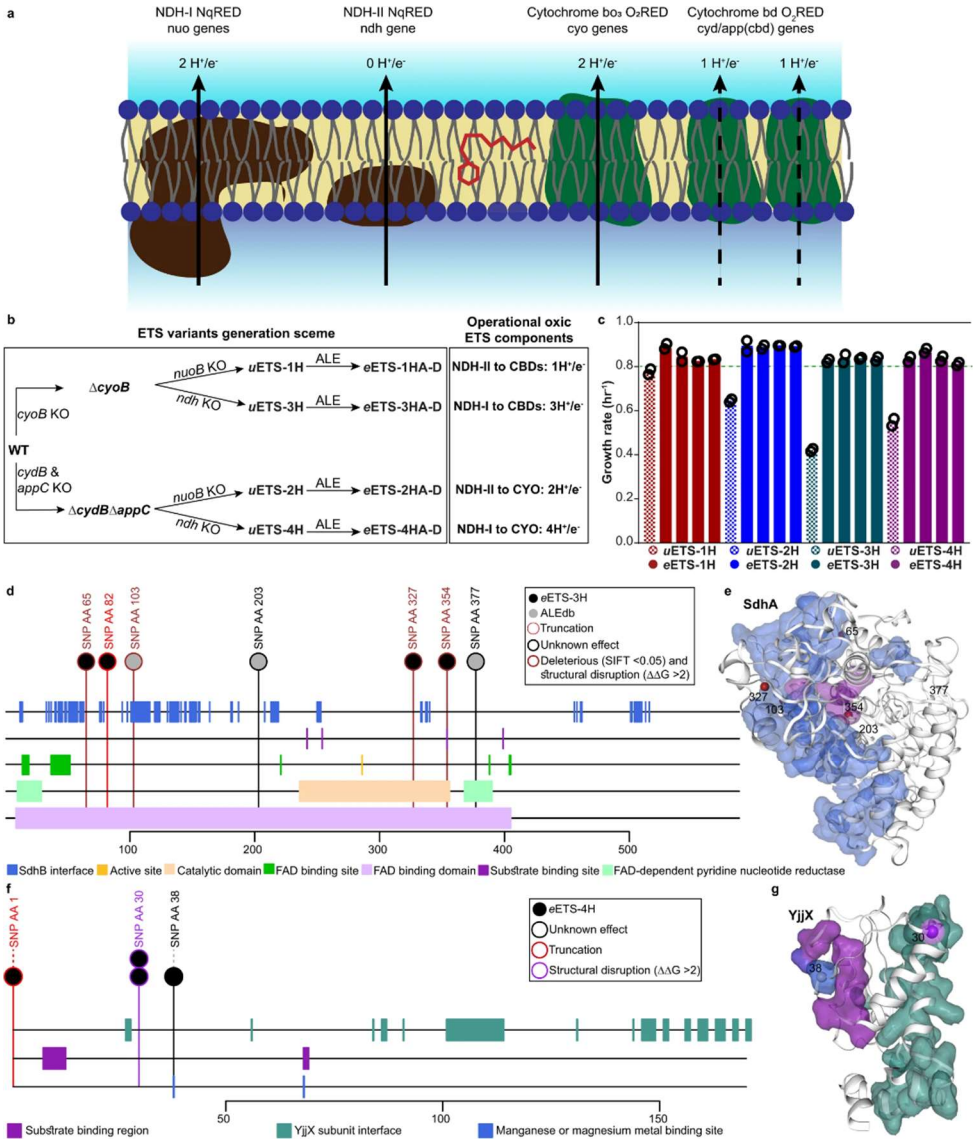


Figure 3.1: Generation and evolution of unbranched ETS variants.

(A) Schematic showing the respiratory enzymes involved in the flow of electrons from NADH (donor) to oxygen (acceptor). NDH-I and NDH-II are the proton pumping and non-pumping NADH: quinone oxidoreductase, respectively. Dashed arrows for CBDs represent the vectorial mode of PMF generation. **(B)** Scheme for generating ETS variants translocating 1, 2, 3, or 4 proton(s) per electron. *u*ETS is the unevolved strain and *e*ETS is the evolved strain. A–D are the four independently evolved lineages of each strain. **(C)** Growth rates of ETS variants before and after ALE. **(D)**, **(F)** Predictive mechanistic interpretation of the impact of mutations observed in the evolved strains of (**D**) ETS-3H (*sdhA*) and (**F**) ETS-4H (*yjjX*) mutations on the structure and function of the protein. Mutations displayed are those from this study and other ALE experiments in ALEdb seen to mutate these genes (76). Horizontal tracks display the reported features associated with the region of the protein. The mutations collected from ALEdb refer to experiments from the following set, respectively (77–79). Protein structures showing the amino acid residues mutated in (**E**) SdhA and (**G**) YjjX.

Based on these characteristics, we designed four ETS variants with unbranched electron flows representing all alternate oxic respiratory routes translocating 1, 2, 3, or 4 proton(s) per electron (designated as ETS-nH, with $n = 1, 2, 3, 4$). The designs of the four ETS variants are illustrated in **Figure 3.1b**.

Next, we analyzed their growth phenotype (**Figure 1c**). Interestingly, the unevolved variants (called *u*ETS) showed different growth rates that had no clear association with their H^+/e^- value. While the loss of activity of NDH-I showed a lesser growth rate retardation, the deletion of NDH-II significantly compromised the growth rate of the deletion strains.

To allow the ETS variants to overcome the growth defects resulting from gene deletions, we performed ALE with four independent replicates of each variant in an oxic environment (**Supplemental Table B.1**) (evolved variants are named *e*ETS-nHm with the replicate evolutionary endpoints indexed as m = A, B, C, D) (80). We evolved all variants until their growth rate plateaued. ETS-1H, 2H & 3H required evolution for approximately 400 generations, while ETS-4H required approximately 700 generations. In spite of the different number of protons pumped per electron, all four ETS variants evolved to a similar optimized growth rate in replicate evolutions (~0.85 h^{-1}) (**Figure 3.1c**).

Next, we sought to determine the acquired mutations that enabled adaptation to a higher growth rate for all ETS variants. We performed whole-genome sequencing of each strain and used a comprehensive database of mutations from ALE experiments (aledb.org (76)) to interpret the potential impact of the identified mutations. The mutation calling revealed only a few genetic changes in the evolved strains except for *e*ETS-4HC which acquired 15 genetic changes. The higher number of mutations in *e*ETS-4HC could be due to the mutated DNA mismatch repair enzyme *mutS* in this strain (81). Every ETS variant acquired mutations responsible for enabling faster growth on M9 minimal medium (Supplemental Table B.2) (77,82–84). An intergenic

mutation between *pyrE* and *rph* has been reported to alleviate pyrimidine pseudo-auxotrophy resulting in a faster growth rate. RNA polymerase subunit mutations are proposed to favor a higher growth rate by accelerating the transcriptional processes. Another common mutation reported to support a faster growth rate is in the intergenic region between *hns* and *tdk*. This mutation is expected to downregulate several stress response pathways and shift resources to support growth. *u*ETS-1H carried the *pyrE-rph* intergenic mutation, which explains the relatively faster initial growth rate of this strain.

Besides mutations responsible for acclimatization to media, *u*ETS-3H and *u*ETS-4H acquired a common gene-related mutation in all four independently evolved lineages. This mutational convergence simplified the otherwise difficult task of establishing the genotype-phenotype relationship (85–87).

All four evolved replicates of *u*ETS-3H acquired point mutations in *sdhA*, the catalytic subunit of succinate dehydrogenase (Supplemental Table B.2). *e*ETS-3HB acquired a point mutation that brings in a premature termination codon in the *sdhA* open reading frame, suggesting a loss of functional enzyme. We explored the potential impact of other mutations by investigating whether the SNPs could affect the protein's function based on amino acid properties and sequence homology (SIFT) (88) or structural stability (ΔΔG) (89). Almost all mutations in *sdhA* were either in or near interface surfaces and seem to be working to disrupt its functionality by either disrupting a substrate-binding site or causing a structural-functional perturbation (**Figure 3.1d, Figure 3.1e**). Notably, the deletion of another subunit of this enzyme, *sdhC*, has been reported to increase the biomass yield in an oxic environment (90). *u*ETS-3H appeared to adopt a similar metabolic route to increase its growth rate.

All four replicates of *u*ETS-4H acquired mutations in an inadequately characterized gene, *yjjX* (**Supplemental Table B.2**). The structural and biochemical evidence suggests that YjjX,

an inosine/xanthosine triphosphatase, may be involved in the mitigation of the deleterious impact of oxidative stress by preventing the accumulation of altered nucleotides (91). Also, the physical association of YjjX with the elongation factor suggests a negative impact on the translational rate. The STRING-based protein-protein interaction predicted the association of YjjX with glycolytic and ATP biosynthetic processes (92). Interestingly, *e*ETS-4HA replaced the start codon, ATG, with ATA. Apart from replacing methionine with isoleucine, this substitution potentially diminishes the expression of this protein (93). A similar disruptive impact is expected from other *yjjX* mutations (**Figure 3.1f**, **Figure 3.1g**). The cysteine to tyrosine substitution at amino acid residue 30 was predicted to destabilize the structure as it lies just beside a subunit interface residue, and a charge reversion due to the glutamate to lysine substitution at amino acid residue 38 targets the metal-binding site. Thus, it appears that *e*ETS-4H is attempting to prevent translational halting to achieve a higher growth rate.

Since the restoration of the evolved variants to the same growth rate cannot be deciphered from genetic changes alone, we took a broader systems view to understand the underlying metabolic perturbations. We examined how the evolved variants rewired the fluxes through the major metabolic pathways that couple to the ETS. We generated RNA sequencing and metabolite profiling data for all the strains and performed targeted and systems level analyses. We observed a high transcriptional correlation among the evolved replicates (Spearman's rank correlation coefficient >0.75) of each variant, but the correlation between pre-and post-evolved variants was lower (**Figure 3.2a**). Notably, consistent genetic and transcriptomic changes supported a common evolutionary trajectory for the replicates of each variant.

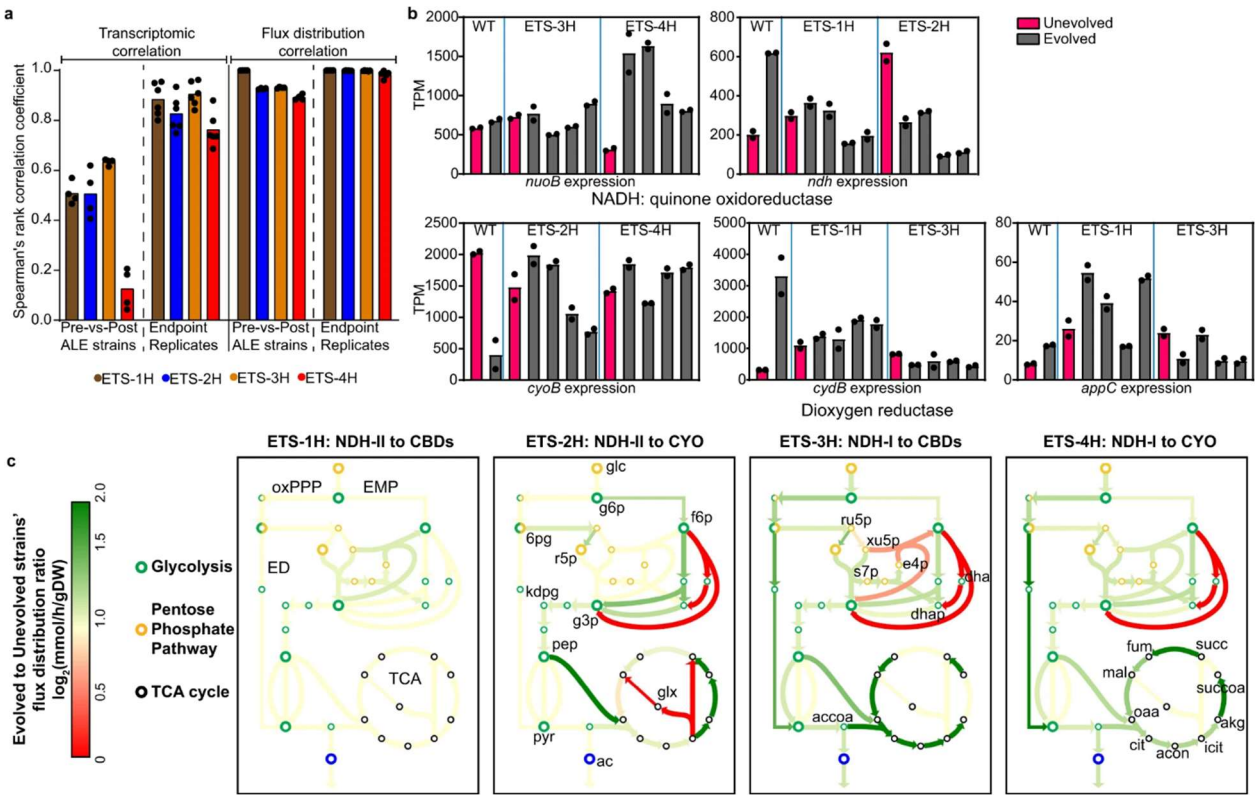


Figure 3.2: Metabolic rewiring supporting growth rate optimization in the ETS variants.

(A) Transcriptional and metabolic flux distribution correlations between evolved replicates and unevolved strains. Correlation of all four evolved replicates of each ETS variant has been calculated with corresponding unevolved strain and has been presented as Pre-vs.-Post ALE strains correlation. Correlation of all four evolved replicates of each ETS variant has been calculated among each other and has been presented as Endpoint Replicates correlation. **(B)** Expression changes in the alternate NqRED and O_2RED in the unbranched ETS variants. **(C)** Computed metabolic flux maps depicting the central metabolism in the evolved ETS variants as compared to respective unevolved ETS variants. Key metabolites are indicated in the figure as follows: glc glucose, g6p D-glucose-6-phosphate, f6p D-fructose-6-phosphate, 6 pg 6-phospho-D-gluconate, ru5p D-ribulose-5-phosphate, r5p α-D-ribose 5-phosphate, xu5p D-xylulose-5-phosphate, s7p sedoheptulose-7-phosphate, e4p D-erythrose-4-phosphate, dha dihydroxyacetone, dhap dihydroxyacetone-phosphate, kdpg 2-keto-3-deoxy-6-phosphogluconate, g3p glyceraldehyde-3-phosphate, pep phosphoenolpyruvate, pyr pyruvate, ac acetate, accoa acetyl-CoA, oaa oxaloacetate, cit citrate, acon cis-aconitate, icit isocitrate, akg 2-oxoglutarate, succoa succinyl-CoA, succ succinate, fum fumarate, mal malate. [oxPPP oxidative pentose phosphate pathway, EMP Embden-Meyerhof-Parnas pathway, ED Entner–Doudoroff pathway, TCA Tricarboxylic acid cycle].

Bacterial physiology displays a remarkable compensatory potential facilitated by altered metabolic flux states resulting from genetic and transcriptomic changes (90). Therefore, we examined if the surrogate NqRED or O_2RED compensated for the loss of function resulting from deleted ETS enzymes (**Figure 3.2b**). There was no clear compensatory trend in the strains with

unbranched ETS except for ETS-4H. ETS-4H increased the expression of NDH-I while increasing or maintaining the expression of CYO after evolution. Surprisingly, the compensatory upregulation of *ndh* in *u*ETS-2H was lost after evolution to a higher growth rate.

Since RNA expression levels may not correlate with metabolic fluxes due to differential translation efficiency and different enzyme catalytic turnover rates, we performed a metabolic flux distribution analysis. To obtain the metabolic flux map, we measured the medium exchange rates of the major metabolites related to respiratory metabolism (**Supplemental Table B.3**). We used both the metabolite exchange rates and transcriptomic data as constraints to simulate the flux through the pathways of the central carbon metabolism using a genome-scale model of metabolism and protein expression (ME-model) (94). We observed a high correlation in the metabolic flux distributions of the four evolved replicates of each strain, further supporting a similar evolutionary pathway followed by replicates of each variant (**Figure 3.2a**).

To more deeply understand the different metabolic states exhibited by the evolved variants, we examined the variations in their computed proteome allocation using the solutions from the phenotypic and transcriptomic constrained ME-models. We observed a clear distinction between strains with alternate NqRED for the preferred glycolytic pathway (**Figure 3.2c**). NDH-I has approximately 10-times higher molecular mass as compared to NDH-II (95,96). Therefore, despite its PMF generation potential, NDH-I is a less preferred dehydrogenase during oxic respiration to achieve faster growth (73). The non-proton pumping high turnover dehydrogenase, NDH-II, is better suited to relieve the growth bottleneck that may arise due to excess built-up of PMF while allowing the operation of oxic ETS (73,97).

The finite resource carrying capacity of a cell creates metabolic tradeoffs on how to partition the proteome to support metabolic pathways best suited for a given growth condition. With an approximately 3.5-fold higher protein cost, the Embden–Meyerhoff–Parnass (EMP)

pathway consumes a larger proportion of proteome as compared to the Entner–Doudoroff (ED) pathway (98). However, the higher ATP yield of the EMP pathway alludes to a potential tradeoff between the two glycolytic pathways for optimizing ATP production while maintaining a growth-supporting proteome (99). The ETS-3H and ETS-4H strains forced to respire using larger NqRED (NDH-I) increased the flux through the proteome conservative ED pathway. Thus, we observed a compensatory selection of the preferred pathway to achieve a balanced proteome.

Interestingly, while strains with *nuoB* deletion (ETS-1H and ETS-2H) increased metabolic flux through complex II of ETS, ETS-3H appeared to minimize the flux through complex II. Notably, *e*ETS-3H lacks *ndh* and acquired a mutation in the gene *sdhA* which codes for a complex II subunit. However, ETS-4H, which also lacks the *ndh* gene, increased the flux through complex II, albeit at a lower level compared to ETS-1H and ETS-2H.

Thus, metabolic plasticity (reflected in metabolic rewiring and associated proteome allocation) allows for redundancy in the *e*ETS variants while supporting the same growth rate. Knowledge of this metabolic plasticity motivated the examination of the overall bioenergetics-state of the evolved ETS variants to fully understand the basis for the evolution to the same growth rate. We have earlier defined an approach to classify the *E. coli* phenotypes into aero-types, which is a quantitative fitness descriptor based on cellular respiratory behavior and proteome allocation (72). The stratification of aero-types is based on the multimodal distribution of the fraction of total ATP produced through ATP synthase which is modulated through the discrete usage of ETS enzymes. We have reported a non-uniform distribution of phenotypic growth data in the rate-yield plane that can be approximately segregated in different aero-types based on sampling simulations. Here we used aero-types to examine the fitness distribution of ETS variants.

We observed that ETS-1H, ETS-2H, and ETS-3H did not show a major shift in their biomass yield during evolution and thus preserved their respective aero-types (**Figure 3.3a**). The

evolutionary optimization of growth rate appears to be largely driven by rewiring central carbon metabolism while oxidative energy metabolism is conserved. ETS-4H jumped from a lower to a higher aero-type after evolution, suggesting an increase in oxic metabolism. The ETS-4H variant has the highest PMF generation capacity. Its aero-type shift to higher classes occurred only after adaptive evolution.

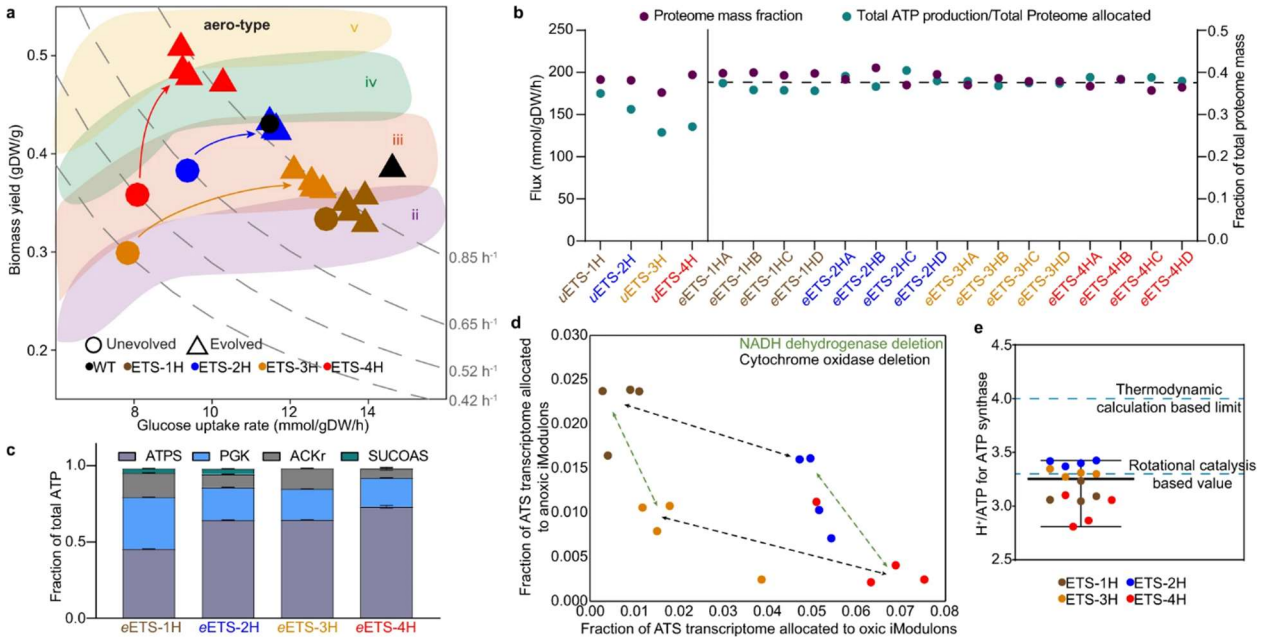


Figure 3.3: Systems-level examination of ETS variants.

(A) Aero-type classification of the ETS variants. Broken lines on the aero-type plot show growth rate isoclines. **(B)** ME-model-based examination of the ATP production (left y-axis) and proteome allocation (right y-axis) in the ETS variants. The ATP produced per ATS proteome is approximately the same. **(C)** Contributions of different ATP-producing reactions towards total ATP production. ATP production by (i) ATP synthase (ATPS) in oxidative phosphorylation, (ii) acetate kinase (ACKr) in mixed acid fermentation, (iii) succinyl-CoA synthetase (SUCOAS) in the TCA cycle, and (iv) phosphoglycerate kinase (PGK) in the glycolysis pathway is shown in the histogram. The mean and standard deviation values are calculated using four independently evolved replicates of each strain. **(D)** Tradeoffs in the expression levels of genes of iModulons associated with anoxic (y-axis) and oxic (x-axis) energetics underlie the rewiring of the ATS to allow all variants to achieve approximately the same growth rate. The lowest aerotype (ETS-1H) has high anoxic/low oxic gene expression while the highest aerotype (ETS-4) exhibits the opposite. The gene composition of the iModulons is shown in **Supplemental Table B.4**. The outliers of the replicates for an ETS variant are reflections of the differences in their genotypes (**Supplemental Figure B.2**). **(E)** Estimation of the number of protons required for the phosphorylation of ADP by ATP synthase (proton-to-ATP ratio) using the ME-model and the experimental data. Individual values of four independently evolved replicates of each ETS variant have been shown on the plot and corresponding median and range of values have been presented.

The clustering of each evolved ETS variant along the same growth rate isocline (**Figure 3.3a**) indicated global remodeling of the energy metabolic network to produce similar growth-supporting bioenergetics. We thus defined a larger respiratory system, called the Aero-Type System (ATS), consisting of oxidative phosphorylation, glycolysis, pyruvate metabolism, the TCA cycle, and the Pentose Phosphate pathway, that together define the overall state of oxic energy metabolism (**Supplemental Figure B.3**). The total proteome allocated to the ATS was very similar in each *e*ETS variant, and the total ATP output of each proteome expressed was almost constant (**Figure 3.3b**). Thus, the composition of the ATS was malleable and able to provide the same supply of ATP, allowing similar growth rates for all *e*ETS variants. We also observed a trend in the metabolic location of ATP production across the variants, where the relative contribution of oxidative phosphorylation was highest for *e*ETS-4H and lowest for *e*ETS-1H (**Figure 3.3c**). Accordingly, an inverse trend was observed for glycolytic and fermentative ATP production.

We next examined the transcriptome to identify the tradeoffs in gene expression that enabled the different metabolic states. We applied a blind source signal separation algorithm, called independent component analysis (ICA) (13), to examine differential partitioning of the transcriptome of the 209 ATS genes. ICA decomposed the ATS transcriptome into independently modulated sets of genes (called iModulons). The activities of several iModulons showed a clear association with the aero-type of the ETS variants (**Supplemental Figure B.4).** iModulons consisting of genes associated with oxic respiration showed a positive correlation with aero-type status (iModulons 8, 13, and b2287), and those constituted by anoxic and/or metabolic genes showed a negative correlation (iModulons 7, 9, 10, 16, and b3366) (**Supplemental Figure B.2**). Thus, an oxic-anoxic transcriptomic tradeoff enabled the four ETS variants to maintain similar ATP production capacity (**Figure 3.3d**).

The direct measurement of the number of protons translocated through ATP synthase to produce one molecule of ATP (H^+/ATP) is technically challenging and, therefore, it is still an area of active research (100). The rotational catalysis-based calculation suggests the H^+/ATP value to be 3.3, due to the symmetry mismatch between the F_o and F_1 complexes of ATP synthase: threefold symmetry of α3β3 in F1 and tenfold symmetry of the c-ring in F_o (101,102). The proton-to-ATP ratio may vary depending upon any change in the number of c-subunits and this modulation allows tailoring to meet the bioenergetic demand of various organisms (103). The H^+/ATP value derived using a synthetically reconstituted membrane system was found to be 4 (104). With our comprehensive definition of the state of the ATS amongst the variants, we could address the issue of ATP synthase proton-to-ATP ratio. We used data generated on the variants to computationally estimate the most likely proton-to-ATP ratio for *E. coli* ATP synthase (100). We constrained the ME-model using the observed metabolic exchange rates and gene expression data and optimized for the H^+/ATP value of ATP synthase that produces the experimentally estimated growth rates of the variants. The ME-model calculates the median value of the H^+/ATP to be 3.25, a value close to 3.3 supporting the rotational catalysis hypothesis (**Figure 3.3e**). Notably, while 10 is the preferred number of *c* subunits in the *E. coli* F_o motor of ATP synthase, the number of subunits can vary, which will change the H^+/ATP value (105–108).

Taken together, our results lead to an expanded definition of oxic respiration beyond the conventional ETS, which involves an electron transport chain to create PMF, that then drives the ATP synthase. Here, we define the Aero-Type System that encompasses the ETS and coupled metabolic pathways (**Supplemental Figure. B.3**). The ATS is composed of 209 genes. The ATS represents about 38% proteome allocation in all evolved variants. A decrease in the ETS energetic efficiency (often measured in terms of the P/O ratio) can be balanced by increased flux through the coupled metabolic pathways. This balance is governed by the cost of protein synthesis.

Remarkably, the overall proteome allocation to the ATS is similar in the evolved variants and generates the same amount of ATP, enabling them to achieve the same growth rate. The different ways in which the ATS is balanced underlies its plasticity and represents a demonstration of the key systems biology concept of alternate optimal states. These alternate states have a different combination of proton pumping efficiency, complementary metabolic rewiring achieved through tradeoffs in the composition of the transcriptome, and concomitant efficiency of proteome allocation, but enable the same overall cellular function. The cytoplasmic-periplasmic adaptive nexus that the ATS represents thus illustrates the deep plasticity inherent in achieving balanced energetic systems to match metabolic needs in different environmental niches.

3.3 Methods

3.3.1 Examining PRECISE 2.0 for expression levels of respiratory enzymes

PRECISE 2.0 is a compendium of high-quality RNA-seq for *E. coli* K-12 (14). It contains 815 RNA-seq datasets of samples with different genetic changes or varied growth conditions. We examined the expression of respiratory dehydrogenases and reductases in the entire dataset. For intelligible purposes, we plotted the expression levels in samples that are directly or indirectly associated with energy metabolism. The expression levels shown are the median value across replicates for a sample.

3.3.2 Strain generation and adaptive laboratory evolution

E. coli K-12 MG1655 (ATCC 700926) was used as the wild-type strain. P1 phage transduction method was used to generate the knockout strains (109), and strains from the Keio collection were used as a donor for the gene knockout cassettes (110). *u*ETS-1H and *u*ETS-3H were generated and used for validation purposes in an earlier study (72). *u*ETS-2H and *u*ETS-4H were generated here and all four ETS variants were evolved for this study.

ALE was performed using 4 independent replicates of each ETS variant. Cultures were serially propagated on M9 minimal medium with 4 g/L glucose at 37 °C and well-mixed for proper aeration using an automated system that passed the cultures to fresh flasks once they had reached an A_{600} of 0.3 (Tecan Sunrise plate reader, equivalent to an A_{600} of ~1 on a traditional spectrophotometer with a 1 cm path length). Cultures were always maintained in excess nutrient conditions assessed by non-tapering exponential growth. The evolution was performed for a sufficient time interval to allow the cells to reach their fitness plateau.

3.3.3 Prediction of the effect of amino acid substitutions

The ALE mutation datasets supporting the conclusions of this article is available in the following open-access archive repository: https://doi.org/10.5281/zenodo.5431595. These datasets are also available in the ALEdb database (76).

Mutated DNA sequence data processing was performed using Python 3. The mutations from ALEdb are described according to their experiment, evolution replicate, sample, and technical replicate. Some evolutions include midpoint samples that could inflate the frequency a mutation is observed. Unique ALE mutations were therefore only considered once per ALE. Starting strain mutations and hypermutator samples were filtered out of the ALE experiment mutation datasets according to their publications. Mutation needle plots were generated using the trackViewer R software package (111). The visualizations for the 3D protein structures were generated using the *NGL* software package (112). The software implementation of these actions is available in the following open-access archive repository: https://doi.org/10.5281/zenodo.5431595.

Mutation effects were predicted according to multiple methods. Truncations were predicted according to the potential effect of mutations on the function of start codons and their potential to introduce a premature stop codon. The predicted deleterious effects of SNPs were assumed according to significant SIFT (sorting intolerant from tolerant) scores (SIFT score < 0.05) (88). The

predicted structural destabilization effects of SNPs were assumed according to predicted significant ΔΔG scores (ΔΔG > 2) (89). SIFT and ΔΔG scores were acquired from *Mutfunc* (113). Functional annotations were acquired from *UniProt* (11) and *Mutfunc*.

3.3.4 DNA sequencing and RNA sequencing

A clone from the endpoints of evolved strains was picked for DNA sequencing and RNA sequencing. The strains were grown in an M9 minimal medium supplemented with 4 g/l glucose. Total DNA was sampled from an overnight grown culture and total RNA was sampled from a culture at an A_{600} ~0.6. Nucleic acid isolation, library preparation, and subsequent analysis were performed as previously described (114). Briefly, genomic DNA was isolated using a Nucleospin Tissue kit including treatment with RNase A. Resequencing libraries were prepared following the manufacturer's protocol using Nextera XT kit. RNA was isolated using the Qiagen RNeasy Mini Kit following suggested protocol. Ribosomal RNA was removed using Illumina Ribo-zero kit and a KAPA Stranded RNA-Seq Kit (Kapa Biosystems KK8401) was used to prepare sequencing libraries. Sequencing was performed on an Illumina HiSeq and/or NextSeq.

3.3.5 Phenotype characterization

Phenotype characterization was performed using two independent biological replicates. Samples for the substrate uptake and secretion rate were collected at regular intervals and filtered using a 0.22 μm filter (PVDF, Millipore). The measurements were performed using refractive index detection by HPLC (Agilent 12600 Infinity) with a Bio-Rad Aminex HPX87-H ion exclusion column. The HPLC method was the following: injection volume of 10 μL and 5 mM H_2SO_4 mobile phase set to a flow rate and temperature of 0.5 mL/min and 45 °C, respectively. The phenotype dataset was used for the aero-type classification of the strains as described previously (72).

3.3.6 Metabolic flux mapping and estimation of H^+/ATP value for ATP synthase

Flux mapping was done as previously described using a genome-scale model of metabolism and protein expression (34). The same FoldME model was used for estimating the H^+/ATP value for ATP synthase within each ETS variant and replicates. The model was constrained with phenotypic data (glucose uptake rate, acetate production rate) and expression data was layered on using the same methods used for the flux mapping (34). In addition to these constraints, the necessary ETS genes for each variant were knocked out. Proton pumping ratios from 2.5 to 4.5 were sampled by changing the stoichiometry of the ATPS4rpp reaction in the ME-model, and then the proton pumping ratio was optimized so that the model produced a biomass dilution rate that matched the experimentally determined growth rate.

3.3.7 ATS proteome allocation calculation

The same FoldME model was used for the proteome allocation calculation as the flux mapping and ATP synthase estimation calculations. The model was constrained with phenotypic data (glucose uptake rate, acetate production rate, growth rate) and expression data was layered on using the same methods used for the flux mapping. Solutions from the fully constrained ME-models were then used for calculating proteome allocation. Total proteome allocation for each strain was calculated as follows:

$$Total\ Proteome\ Allocation = \sum_i mw_i * V_i^{translation}$$

Where mw_i and $V_i^{translation}$ represents the molecular weight and translation flux of the *i*th protein in the model. Total proteome allocated to the ATS was calculated as follows:

$$Proteome\ Allocated\ to\ ATS = \sum_i mw_i * V_i^{translation}$$

where mw_i and $V_i^{translation}$ represents the molecular weight and translation flux of the *i*th protein in the ATS (209 genes total). The list of 209 ATS genes was generated based on Clusters of Orthologous Groups (COG) and Gene Ontology (GO) categories to include as many relevant genes

as possible to represent pathways involved in ATP production, then filtered to remove genes that are never expressed in the multiple model simulations. Mass fraction of proteome allocation to the ATS was calculated as a ratio of the two values for each strain.

Calculation of the total ATP produced by the ATS used the same fully constrained ME-model. A list of all metabolic reactions associated with ATS genes was curated. Reactions that consumed or produced ATP were noted and the stoichiometric coefficient associated with ATP was used as a modifier for calculating the total ATP production as follows in **Table 3.1**.

$$Total\ ATP\ Production = \sum_i c_i * V_i^{metabolic}$$

where c_i and $V_i^{metabolic}$ represents the ATP stoichiometric coefficient and the metabolic flux of the *i*th ATS associated reaction in the **Table 3.1**.

Table 3.1: Reactions that consume or produce ATP and corresponding stoichiometric coefficient.

Reaction[a]	**ATP coefficient**
PPK	−1
PPK2	−1
ATPS4rpp	1
PFK_2	−1
HEX1	−1
PYK	1
PFK	−1
PPS	−1
PGK	−1
GLGC	−1
PFK_3	−1
GART	−1
PPAKr	1
ACCOAL	−1
ACS	−1
ACKr	−1
SUCOAS	−1

[a]Reaction IDs with corresponding reactions in the BiGG Database (bigg.ucsd.edu).

Total ATP Production/Total Proteome Allocated was calculated as a ratio of the total ATP production to the mass fraction of proteome allocated to the ATS for each strain.

3.3.8 ATS transcriptome ICA decomposition

Independent component analysis was performed on an RNA-seq dataset with steps described in (14). The only genes included in the dataset were those contained in the list of 209 ATS genes. The dataset consisted of all unevolved strains, *u*ETS-1H through 4H, and all evolved replicates *e*ETS-1HA through *e*ETS-4HD. Additionally, the unevolved and evolved wild-type strains were included with the former being used as a reference to center the data. The final and resulting dataset that was used for ICA contained 209 genes by 22 conditions.

3.3.9 Data and code availability

Resequencing and expression profiling data that support the findings of this study can be accessed from NCBI Sequence Read Archive accession number PRJNA835443 and Gene Expression Omnibus accession number GSE202144 respectively. The PRECISE compendium and all associated data files can be found at https://github.com/SBRG/precise2. Source data for the figures can be found in **Supplemental Tables B.2 and B.3** as well as in the uploaded RNA-seq data.

The software scripts supporting the prediction of mutation effects to the encoding of genes described in this article are available in the following open-access archive repository: https://doi.org/10.5281/zenodo.5431595. All the simulations performed in this manuscript can be reproduced using the FoldME model, which is constructed using the COBRApy toolbox version 0.5.11 for constraint-based modeling and its extension for ME-models, COBRAme version 0.0.9, ECOLIme version 0.0.9, and solveME, all publicly available on Github (https://github.com/SBRG/ME-script, https://github.com/SBRG/ecolime, https://github.com/SBRG/solvemepy). Custom code for

constraining and solving ME-models can be found at https://github.com/SBRG/ME-script. GraphPad Prism version 9.2.0 was used for generating the plots.

3.4 Acknowledgments

This work was funded by the Novo Nordisk Foundation Grant Numbers NNF10CC1016517 and NNF20CC0035580 and National Institutes of Health Grant R01GM057089. We would like to thank Marc Abrams (Systems Biology Research Group, University of California San Diego) for assistance with paper editing.

A.A., A.F., and B.O.P. designed the study. A.A., C.O., and R.S. performed the experiments. A.A., A.P., K.C., P.P., C.L., and B.O.P. analyzed the data. A.A. and B.O.P. wrote the paper, with input from all co-authors.

This section of Chapter 3, in part, is a reprint of the material as it appears in *Nature Communications*:

Anand A, **Patel A**, Chen K, Olson CA, Phaneuf PV, Lamoureux C, et al. Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity. Nat Commun. 2022 June 27;13(1):3682.

The dissertation author was the second author.

3.5 Background

Energy conservation pathways and associated metabolic flexibility are critical for bacterial physiology (115). The need for the indispensable energy carrier molecule, ATP, is met from several metabolic locations in a cell, with oxidative phosphorylation facilitated by ETS being the primary source. This diversity in energy generation pathways allows bacteria to thrive under various environmental conditions by exploiting substrates that differentially feed into various metabolic reactions. Such flux optimization to achieve robust growth physiology requires an interplay among glycolytic, respiratory, and fermentative pathways.

The pathways' interplay occurs at a level beyond the classical regulon. A set of gene groups that are co-regulated in response to specific environmental stimuli or stress conditions may not be regulated by a specific regulatory protein. Such co-regulated gene groups are referred to as stimulons. This framework gained traction in the late 20th century, with early studies focusing on bacterial shock responses and stress adaptations (116,117). However, the enthusiasm surrounding stimulons waned in the early 2000s due to challenges related to clearly defining their boundaries, their overlap with regulons, and the technical limitations of early transcriptomics (118,119). These difficulties and a shift towards systems biology diminished their relevance.

Recent advancements in RNA-Seq, computational modeling, and integrative omics have rekindled interest in stimulons by providing robust frameworks for their identification and analysis. Modern studies have demonstrated the value of stimulons in understanding adaptive responses to oxidative stress, nutrient deprivation, and other environmental challenges (120–122). By leveraging these technologies, researchers have begun to uncover novel aspects of bacterial physiology and reestablish stimulons as valuable tools for metabolic research. The cellular bioenergetic network has intricate regulatory features, which are difficult to discern solely based on the biochemical properties or classical regulon concepts. In a genome-wide screening for ATP synthetic activity, genes believed to function synergistically in supporting oxidative phosphorylation were proposed to be regulated differently (123).

In a recent study, we performed adaptive laboratory evolution of four *Escherichia coli* strains that differed in aerobic ETS composition and, thereby, in the number of proton(s) pumped per electron received from NADH (35). We observed that regardless of the differences in ETS efficiencies, the variants obtained similar growth rates on a minimal medium with glucose as the carbon source by leveraging flexibility in ATP production reactions. We also observed that the Aero-Type System (ATS), a generalized framework of the ETS with additional connecting

bioenergetic pathways that we previously described (35), maintained a constant proteome composition across *E. coli* strains with differing ETS efficiencies, highlighting its role in robust bioenergetic adaptation. However, such bioenergetic equivalency may be limited to respiro-fermentative metabolites like glucose. Despite glucose being a preferred growth-promoting source, *E. coli* needs to survive on varying carbon sources (124). *E. coli* can exploit lipids, short-chain fatty acids, amino acids, and nucleotides to meet carbon and energy requirements. Notably, the microbiota produces abundant SCFAs, such as succinate, in the mammalian gut. The chemical nature of the carbon source and the enzymatic composition of the cell determines the flux balance of the catabolic energy generation pathways and anabolic biomass synthetic pathways. In this study, we explore the systems optimization approaches of the ETS variants when supplied with carbon sources that are metabolically more restrictive.

The defect in ETS would be less pronounced on glucose as substrate-level phosphorylation can offset the ATP deficiency. We, therefore, evolved the ETS variants on a non-fermentable carbon source, succinate, to exert strong selection pressure on ETS functioning (125). We also used a fermentable but energy-poor substrate, glycerol, for a parallel evolution (126). Glucose undergoes complete glycolysis, while glycerol enters at the midpoint of glycolysis, contributing to both glycolysis and gluconeogenesis. In contrast, succinate is primarily directed towards gluconeogenesis (**Figure 3.4a-c**). We then performed multi-scale computational analyses on the growth rate-optimized strains. We observed (a) growth optimality on multiple carbon substrates, (b) conservation of proteome allocation to the bioenergetic pathways of the ATS, irrespective of carbon substrate or evolutionary conditions, (c) similar transcriptomic adjustments across seven independent gene groups in all ETS variants studied, revealing an aerobicity stimulon, (d) genome-scale models integrated with omics data identifying the ATS as central to bioenergetic genotype-phenotype relationships, and (e) while glycerol did not require ETS-specific genetic changes,

succinate-evolved strains acquired compensatory mutations. This study defines the aerobicity stimulon, providing a novel perspective on how *E. coli* optimally utilizes its ATS across phenotypes. Our findings also demonstrate how modern methodologies can illuminate and redefine classical biological frameworks like stimulons.

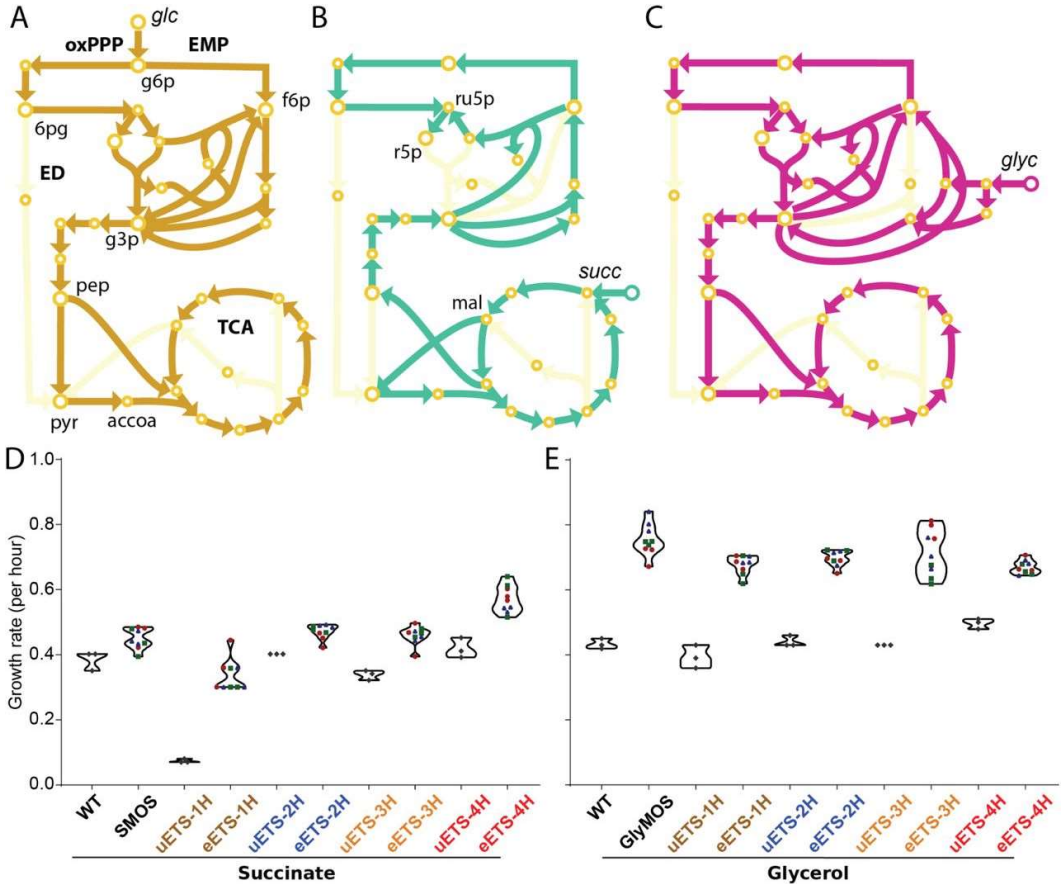


Figure 3.4: Impact of carbon source on ETS variants.

Metabolic flux routing of **(A)** glucose, **(B)** succinate, and **(C)** glycerol in *E. coli* central carbon metabolism. Growth rate of ETS variants on **(D)** succinate and **(E)** glycerol as carbon source before and after growth rate optimization by adaptive laboratory evolution. All laboratory evolutions were performed with three independently evolving lineages. Panels D and E share their y-axis label. The data from the final three flasks of evolution experiments were used to show the growth rates of evolved strains. Different colors represent independently evolved lineages of each strain (Lineage A: red circles, Lineage B: green squares, and Lineage C: blue triangles). [Acronyms-EMP: Embden-Meyerhof-Parnas Pathway, ED: Entner-Doudoroff Pathway, oxPPP: Oxidative Pentose Phosphate Pathway, TCA: Tricarboxylic Acid Cycle, g6p: D-Glucose-6-Phosphate, f6p: D-Fructose-6-Phosphate, 6pg: 6-Phospho-D-Gluconate, g3p: Glyceraldehyde 3-Phosphate, pep: Phosphoenolpyruvate, pyr: Pyruvate, accoa: Acetyl-CoA, ru5p: Ribulose 5-Phosphate, r5p: Ribose 5-Phosphate, mal: Malate, glc: Glucose, succ: Succinate, glyc: Glycerol, WT: Wild Type, SMOS: Succinate Minimal Media Optimized WT Strain, GlyMOS: Glycerol Minimal Media Optimized WT Strain.]

3.6 Results and Discussion

3.6.1 ETS variants exhibit growth optimization and recover on multiple substrates

The ETS in *E. coli* consists of multiple dehydrogenases and reductases, allowing the bacterium to adapt to aerobic and anaerobic environments (73). Previous studies have demonstrated a regulatory hierarchy in *E. coli*, where oxygen represses anaerobic pathways, and nitrate further suppresses alternative electron acceptors (71,73). Despite these controls, the co-expression of multiple respiratory pathways has been shown to enhance flexibility in electron transport (71). Interestingly, multiple electron flow routes exist for the same electron donor-acceptor pair-NADH to oxygen. Despite the same starting and ending metabolic fate, electrons moving through these alternate routes interact with different redox enzymes. Thus, the alternate routes have distinct proton motive force generation abilities. This allowed us to genetically engineer four distinct ETS variants translocating 1, 2, 3, or 4 protons per electron (designated ETS-nH, where n = 1, 2, 3, or 4) (35). Building on our prior work with glucose as a carbon source (**Figure 3.4a**), we extended this framework to investigate the condition-specific performance of the ETS variants under succinate (**Figure 3.4b**) and glycerol (**Figure 3.4c**) metabolism—two carbon sources with distinct metabolic and bioenergetic profiles.

Analyzing the growth phenotype of the unevolved strains (called uETS) reveals that the uETS variants grown on succinate exhibited growth rates clearly associated with their H^+/e^- value, with the lowest H^+/e^- strain, uETS-1H, experiencing the most significant growth retardation (**Figure 3.4d**). This growth pattern confirms the metabolic restriction imposed by non-fermentable carbon sources. In contrast, the glycerol strains showed similar growth rates with no apparent association to their H^+/e^- value (**Figure 3.4e**). For a detailed systems biology study, we started by performing adaptive laboratory evolution (ALE) of the ETS variants in the two media types in oxic

conditions to optimize their growth. The evolved variants, designated as eETS-nHm, were indexed by their replicate endpoints (m = A, B, C). The evolution continued until growth rates stabilized. Similar to the glucose evolutions, despite differences in proton-pumping capacities, all four ETS variants on glycerol ultimately converged on a similar optimized growth rate across replicates (**Figure 3.4e**), whereas they acquired different optimized growth rates on succinate (**Figure 3.4e**). ETS-1H, with the lowest ETS efficiency, failed to match the growth rates of the other variants, while ETS-4H acquired the highest growth rate.

Previously, we established a method to categorize *E. coli* phenotypes into aero-types, a quantitative descriptor reflecting cellular respiratory activity and proteome allocation patterns (72). Aero-types are defined by the multimodal distribution of the proportion of total ATP generated via ATP synthase, influenced by the discrete utilization of ETS enzymes, essentially capturing the aerobicity state of the cell. This classification revealed a non-uniform distribution of growth phenotypes within the rate-yield plane, which could be broadly partitioned into distinct aero-types through sampling simulations. In this study, we applied the aero-type framework to assess the fitness distribution across the ETS variants.

In the previous glucose evolutions, the aerobicity of the evolved strains followed a distinct ranking, with ETS-2H and ETS-4H demonstrating higher aerobic capacities while ETS-3H and ETS-1H exhibiting lower. Specifically, the aerobicity increased as follows: ETS-1H < ETS-3H < ETS-2H < ETS-4H (35). Notably, the same hierarchy in aerobicity is preserved when examining the evolved strains grown on succinate (**Figure 3.5a**) and glycerol (**Figure 3.5b**), suggesting that the differences in aerobicity are inherently linked to the specific ETS complexes.

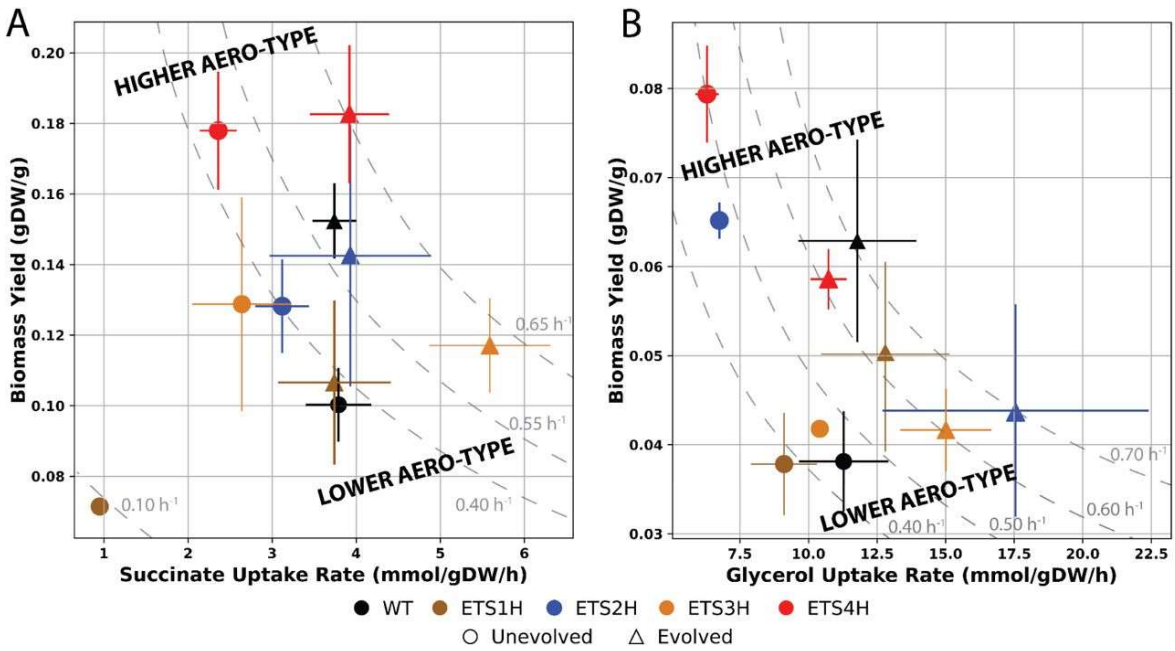


Figure 3.5: Rate yield planes of ETS variants on each carbon source showcase their different aerobic potentials.

Scatter plots of **(A)** succinate and **(B)** glycerol uptake versus biomass yield enable aero-type classification and comparison. Broken gray lines show growth rate isoclines. ETS variants stratify the aero-type spectrum similarly to the strains grown previously on glucose.

3.6.2 Aero-Type System describes convergence in ETS variants' diverse growth optimization

To understand the observed aerobicity patterns, we employed computational models to validate and further explore the bioenergetic behavior of the ETS variants. Specifically, we used metabolite exchange rates, growth rates, and transcriptomics data as constraints for metabolism and expression models (ME-Models) (**Supplemental Table C.6 and C.7**) (6). These genome-scale metabolic models, which incorporate translation and transcription machinery, allow for the analysis of metabolic flux and proteome allocation (127,128). By applying this approach, we generated strain-specific ME-Model solutions for each ETS variant grown on succinate and glycerol.

Analysis of quinone fluxes across the ETS variants revealed a clear association between complex usage and aerobicities (**Figure 3.6a**). The WT strain, which lacks any knockouts, predominantly utilized NDH-I and CYO, a combination that supports high proton-pumping efficiency and is reflective of its high aero-type. In contrast, the ETS variants were constrained by

their engineered knockouts, which dictated their complex usage. For instance, ETS-4H and ETS-2H leveraged NDH-I or NDH-II with CYO, reflecting moderately high aero-types. On the other hand, ETS-3H and ETS-1H, which lacked CYO, relied on bd-type oxidases (CYD/CBD) and menaquinones. This aligns with literature indicating that ubiquinone is the primary quinone utilized under aerobic growth conditions, whereas menaquinone predominates under more anaerobic conditions (129). These observed associations between complex usage and aerobicities highlight how specific combinations of dehydrogenases and oxidases contribute to respiratory efficiency and the bioenergetic state of the cell. These complex usage preferences and their association with aerobicities persist across our previous glucose study, reinforcing the robustness of the observed relationships.

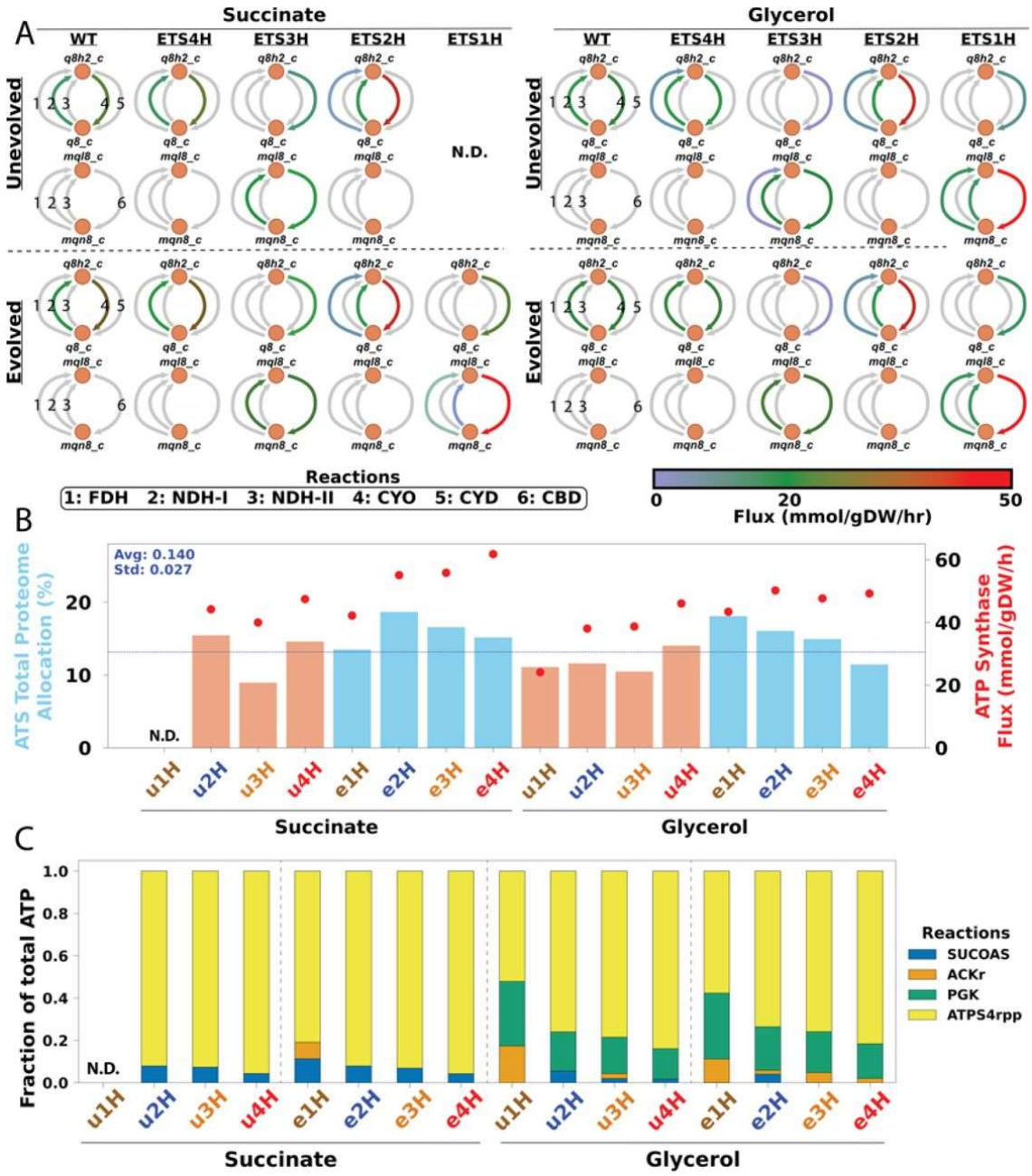


Figure 3.6: Omics-constrained ME models of ETS variants reveal key characteristics of how strains readjust their bioenergetics.

(A) Quinone flux maps for unevolved and evolved ETS variants. Ubiquinones are in the top loop (q8h2_c and q8_c), while menaquinones are in the bottom loop (mql8_c and mqn8_c). Reactions are labeled 1 through 6 and are Formate Dehydrogenase (FDH), NADH Dehydrogenase I (NDH-I), NADH Dehydrogenase II (NDH-II), Cytochrome bo3 (CYO), Cytochrome bd I (CYD), Cytochrome bd II (CBD), respectively. **(B)** Calculated proteome allocation (left y-axis) and flux through ATP Synthase (right y-axis) for all of the ETS variants. **(C)** Stacked bar charts of the fraction of total ATP from ATP-producing reactions for all ETS variants. uXH: unevolved X protons per electron ETS variant, eXH: evolved X protons per electron ETS variant, SUCOAS: Succinyl-CoA synthetase, ACKr: Acetate kinase, PGK: Phosphoglycerate kinase, ATPS4rpp: ATP Synthase.

Interestingly, NDH-I, despite its molecular mass being approximately 10 times greater than NDH-II (95), plays a critical role in optimizing ETS efficiency. While NDH-II enables high-turnover electron transfer and alleviates PMF buildup for rapid growth (97), our findings suggest that NDH-I is more integral for a fully optimized ETS. Similarly, CYO emerged as the most important for ETS efficiency among oxygen reductases, outperforming bd-type reductases CYD and CBD. These preferences for NDH-I and CYO at higher aero-types align with results from a study that modeled 2,200 different phenotype-genotype combinations (72).

Further analysis revealed the activity of formate dehydrogenases (FDHs) varied with aerobicity. In the unevolved glycerol strains, FDH-O was active in the more aerobic ETS-4H and ETS-2H strains, while FDH-N was utilized in the less aerobic ETS-3H and ETS-1H strains. This is consistent with prior studies showing that FDH-O expression is upregulated under aerobic conditions, whereas FDH-N is induced in anaerobic environments (130,131).

Analyzing the proteome allocation to the ATS across all strains reveals a consistent distribution, with the models allocating an average of 0.14 proteome fraction (std = 0.027), regardless of growth rate or ATP production via ATP synthase (**Figure 3.6b**). This finding aligns closely with data from our in-house proteomic expression profiling compendium (n = 162 samples, mean = 0.19, std = 0.027) (**Supplemental Figure C.1**) (132). A similar trend of a consistent proteome was observed in our previous glucose study, further emphasizing the remarkable plasticity of the ATS. Despite this conserved proteome allocation, the strains exhibit different ATP production distributions (**Figure 3.6c**), with higher aero-type strains relying more heavily on ATP synthase and less on fermentative pathways.

3.6.3 Transcriptomic adjustments consistency revealing an aerobicity stimulon

We sought to determine whether specific gene groups could explain the shifts in aerobicities, or aero-types, observed across samples. To investigate this, we utilized iModulon analysis and our comprehensive in-house RNA-seq expression profiling compendium (n = 1,035 samples)(13,14). iModulon analysis employs Independent Component Analysis (ICA), a blind source signal separation algorithm, to decompose RNA-seq expression data into two matrices. The first matrix identifies independently modulated groups of genes, referred to as iModulons (n = 201 iModulons), while the second quantifies the activity levels of each iModulon across all samples in the dataset.

A correlation matrix for all 201 iModulon activities across the 1,035 samples in our compendium was constructed (**Supplemental Figure C.2**). Hierarchical clustering of this matrix, followed by thresholding, identified a small number of iModulons with coordinated activities or significant correlations across all of our data. When we overlaid metadata such as the glucose, succinate, and glycerol ETS specialists, along with anaerobic samples from our compendium, we observed that one iModulon cluster distinctly aligned with shifting aerobicities (**Figure 3.7a**). All scatter plots can be found in **Supplemental Figure C.3**. This cluster, which includes seven seemingly unrelated iModulons, emerged as the largest and most robust cluster in the thresholded correlation matrix.

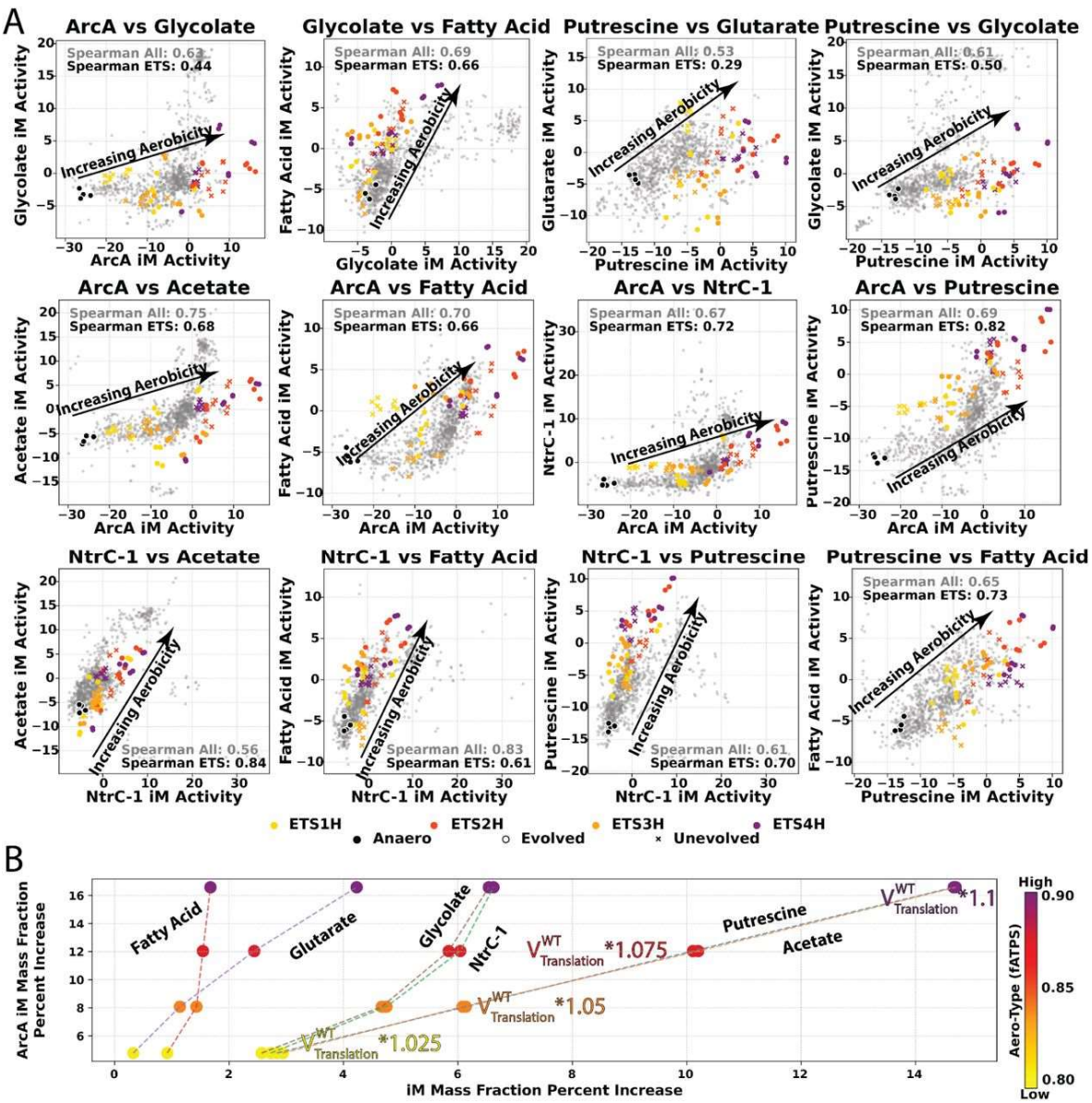


Figure 3.7: Interoperating multi-omics data and computational biology tools for discovery and validation.

(A) Scatter plots of two iModulon's Activities from our in-house transcriptome compendium (12,14) with all glucose, succinate, and glycerol ETS variants labeled. Anaerobic samples are labeled as well to highlight the aerobicity directionality. All other samples in the compendium are colored gray. **(B)** Scatter plots of proteome allocation iModulon mass fractions derived from ME Model simulations where the translation flux of ArcA iModulon genes is iteratively increased, starting with the Wild Type flux. Points are colored by the fraction of total ATP generated through ATP synthase (fATPS), which is a calculation of Aero-Type.

These seven iModulons–ArcA, NtrC-1, Acetate, Fatty Acid, Glutarate, Glycolate, and Putrescine–represent a unified transcriptional response to increasing aerobicities, collectively forming what we term the ***aerobicity stimulon***. Remarkably, the stimulon was consistently

observed across all three substrates studied as well as our entire transcriptomic compendium (n = 1035), highlighting its substrate-independent and genotype/phenotype-agnostic role in coordinating the metabolic and bioenergetic adaptations required for scaling aerobicity. The ArcA iModulon, associated with the transcription factor ArcA, encompasses genes involved in aerobic growth, including those encoding ETS complexes (133). The NtrC-1 iModulon, regulated by NtrC, contains genes from the major arginine degradation pathway in *E. coli*, the arginine N-succinyltransferase (AST) pathway (134). The Acetate iModulon includes genes for acetate uptake and catabolism, while the Fatty Acid iModulon features the fatty acid degradation (*fad*) genes and is regulated by FadR (135). The Glutarate iModulon is associated with the 4-aminobutanoate (GABA) transport/catabolism genes *gabDTP* (136). The Glycolate iModulon comprises the *glcDEFGBA* operon, responsible for glycolate transport and catabolism, and is regulated by the transcription factor GlcC (137). Lastly, the Putrescine iModulon includes genes for putrescine transport and catabolism, organized in the *puuDRCBE* and *puuAP* operons under the regulator PuuR (138).

A ME-Model was used to computationally validate the relationships among the seven iModulons. By constraining a wild-type *E. coli* model with increasing levels of translation flux to ArcA iModulon genes, we observed a corresponding increase in proteome allocation to the other six iModulon gene sets (**Figure 3.7b**). Notably, the Putrescine and Acetate iModulons exhibited coordinated behavior, as did Glycolate and NtrC-1. Furthermore, as these iModulons increased in mass fraction, their aero-type also rose, effectively recapitulating the transcriptomic relationships linked to increasing aerobicities observed in **Figure 3.7a**.

The underlying biological significance of the aerobicity stimulon is further revealed through its integration with metabolic networks. A network map of these iModulons reveals extensive

interconnections and strong links to energy metabolism, revealing their role in effectively scaling aerobicities (**Figure 3.8**).

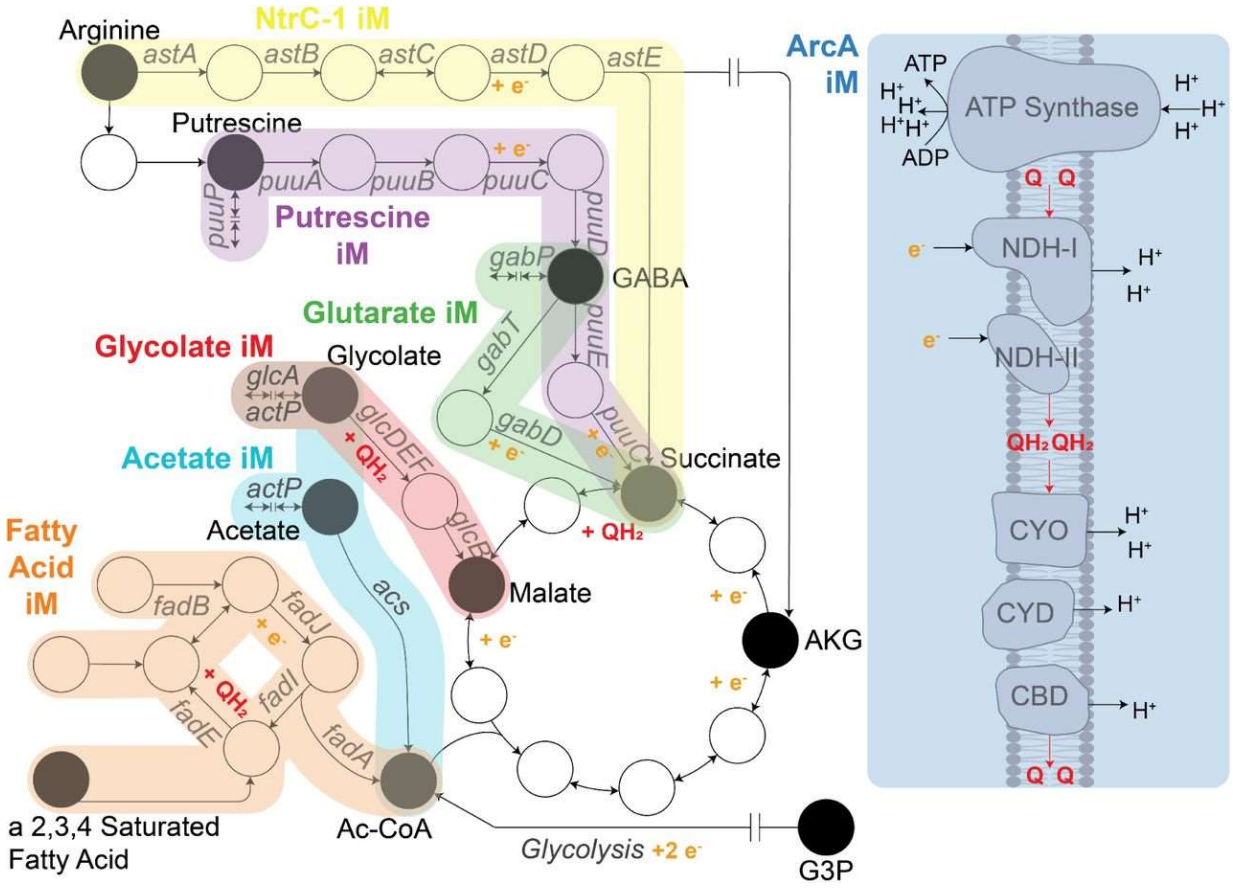


Figure 3.8: Knowledge-enriched transcriptomic data analytics and computational modeling of metabolic fluxes and proteome allocation elucidate an aerobicity stimulon.

A network map of the seven iModulons highlighted in **Figure 3.7** shows their intricate interconnections. These iModulons exhibit a highly coordinated increase in transcriptomic activity as the cell transitions to aerobic conditions. Genes are labeled in gray next to their corresponding arrow. Labeled metabolites are highlighted in black. Quinones are colored in red (QH2: reduced, Q: oxidized), and electron carriers are labeled in orange. NDH-I: NADH Dehydrogenase I, NDH-II: NADH Dehydrogenase II, CYO: Cytochrome bo3, CYD: Cytochrome bd I, CBD: Cytochrome bd II, AKG: 2-oxoglutarate, GABA: 4-aminobutanoate, G3P: Glyceraldehyde 3-Phosphate.

The metabolic network analysis reveals multiple nodes with high connectivity among the iModulons, including GABA, succinate, and acetyl-CoA. Notably, the acetate/glycolate symporter, *actP*, appears in the Acetate iModulon, while a different symporter, *glcA*, is present in the Glycolate iModulon. Both the Putrescine and Glutarate iModulons degrade GABA using succinate-semialdehyde dehydrogenases (*puuC* and *gabD*, respectively); however, *puuC* generates

NADH, whereas *gabD* produces NADPH. The Putrescine and NtrC-1 iModulons connect through agmatine and agmatinase, with Putrescine and Arginine degradation pathways converging to produce succinate. Each iModulon contributes intermediates to the TCA cycle and generates an electron carrier, with the exception of the Acetate iModulon. Furthermore, the Fatty Acid and Glycolate iModulons also produce reduced quinones. Collectively, these intermediates and byproducts enhance the expression of ETS-related genes in the ArcA iModulon, further linking these pathways to aerobic metabolism.

The metabolic complexity and functional interdependencies revealed by this analysis demonstrate that the ETS does not operate in isolation. Instead, it integrates with broader metabolic pathways to enable a systems-level adjustment within the ATS. This alignment of transcriptional responses with respiratory state highlights the critical role of the aerobicity stimulon in bioenergetic optimization.

3.6.4 Genetic changes enabling growth optimization

Phenotypic gains have an underlying genetic basis. We wanted to examine if specific genetic changes supported the growth improvement of the ETS variants on succinate and glycerol as carbon sources. We observed three categories of mutations in the evolved strains: (1) mutations permitting growth-supportive cellular resource optimization, (2) mutations influencing ETS functioning, and (3) mutations influencing specific carbon-source uptake or retention.

There are several well-characterized mutations known to improve the growth rate of *E. coli* on minimal media. The most commonly observed mutations are related to RNA polymerase and Orotate phosphoribosyltransferase (PyrE). These mutations optimize growth-supportive cellular resources either by global transcriptional changes or by alleviating strain-specific metabolic deficiency (77,139). We frequently found fixation of these mutations in the strains evolved on a minimal medium containing either of the carbon sources.

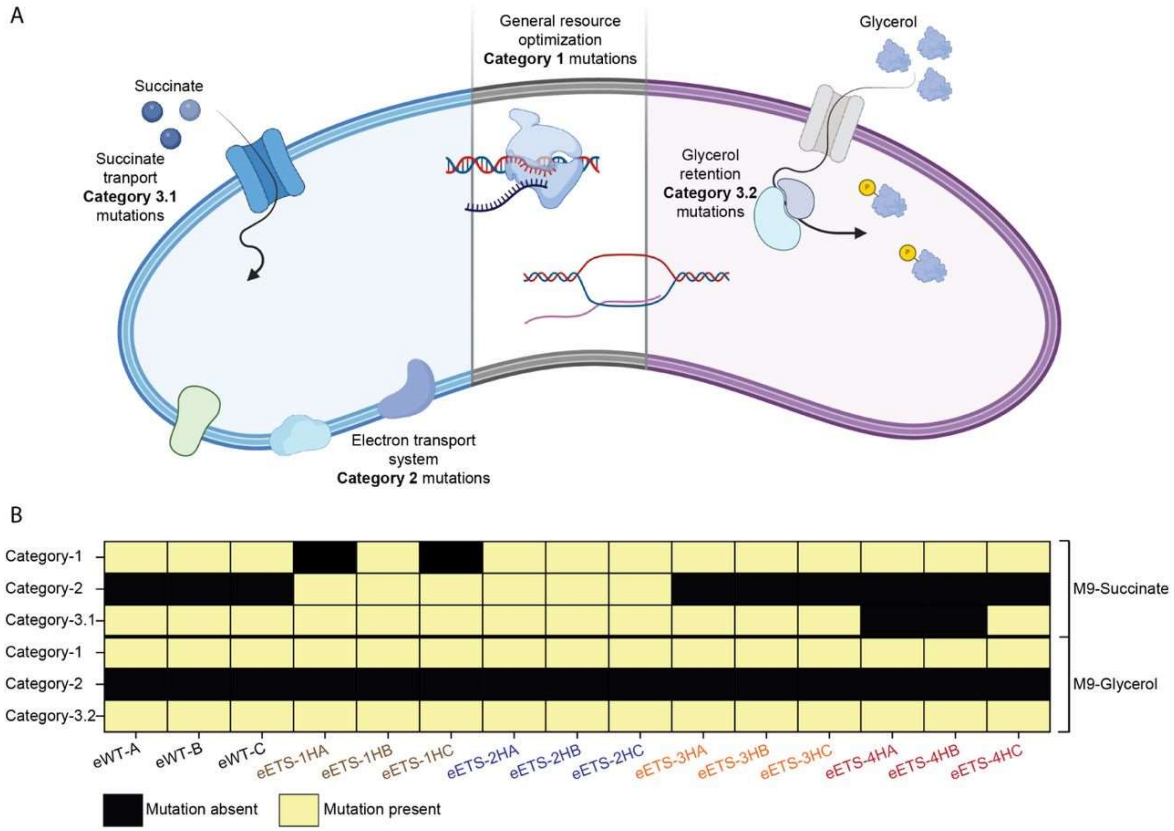


Figure 3.9: Mutations acquired during growth optimization of the ETS variants on succinate and glycerol.

(A) A general scheme for the three major categories of mutations observed. Category 1 mutations are frequently observed during ALE on minimal media and are responsible for supporting growth-supportive cellular resources (77). Category 2 mutations bring about compensatory transcriptional changes in ETS components. Category 3 mutations influence either carbon source uptake (3.1) or cellular retention (3.2). **(B)** Heatmap showing the presence or absence of the three major categories of mutations in strains evolved on minimal medium containing either succinate or glycerol as carbon sources. eWT corresponds to SMOS and GlyMOS for the evolution of WT on succinate and glycerol as carbon sources, respectively. Details of the mutations are listed in **Supplemental Tables C.4 and C.5**.

The need to exploit succinate for metabolic needs resulted in the selection of mutations targeting the C4-dicarboxylate transport system (140). Two common mutations that we observed were in *dctA*, a proton motive force-dependent dicarboxylate transporter, and its transcriptional regulator, *dctR*. Such mutations are reported to improve the succinate transport capacity (141). We also observed mutations in *dcuB* and *dcuD*, other families of succinate transporters. Interestingly, the WT copy of *dcuD* is believed to be a silent gene due to its transcriptional dormancy (140). In a previous study, every *E. coli* lineage evolved on a minimal medium containing glycerol acquired

mutations in the glycerol kinase gene (*glpK*), the rate-limiting enzyme in glycerol metabolism (142). We observed similar mutations in *glpK* in every ETS variant evolved on glycerol as the carbon source.

The two ETS variants, ETS-1H and ETS-2H, evolved on succinate acquired mutations influencing the expression of type II NADH dehydrogenase (Ndh; NDH-II) (143). Notably, these variants have lower ETS efficiency due to the absence of proton-pumping type I NADH dehydrogenase (Nuo; NDH-I). Pyruvate dehydrogenase regulator (PdhR) represses the expression of type II NADH dehydrogenase by binding to the PdhR box upstream of the gene *ndh*. The mutations in this region can alleviate the repression, resulting in the required upregulation of *ndh*. Two eETS-4H strains that evolved on succinate did not acquire category-3.1 mutations, suggesting that these strains' inherent high ETS efficiency alleviates any stress on cellular energy generation pathways. However, the utilization of glycerol as the carbon source did not impose selection pressure on the ETS.

The findings presented here underscore the remarkable adaptability of the bacterial ETS and its associated metabolism when subjected to environmental and genetic perturbations. Engineered strains with unbranched ETS pathways demonstrated their ability to rewire metabolic networks and achieve near-optimal growth rates by leveraging diverse bioenergetic strategies. The observed plasticity of the ATS across multiple substrates highlights its role as a unifying framework for understanding bacterial respiratory flexibility. By enabling *E. coli* to maintain energy homeostasis through alternate optimal states, the ATS reveals how bacteria adapt their metabolic and respiratory strategies to accommodate specific constraints.

The identification of the aerobicity stimulon adds a novel dimension to this understanding, demonstrating how transcriptional networks contribute to bioenergetic optimization. By combining iModulon analysis with genome-scale metabolic and expression models, this study provides a

detailed view of the systems-level adaptations that underpin metabolic resilience. These findings not only deepen our knowledge of bacterial physiology but also establish the ATS as a generalizable descriptor of aerobic bioenergetics, capturing the inherent plasticity and evolutionary versatility of respiration.

3.7 Methods

3.7.1 Strain generation and adaptive laboratory evolution

The unevolved *Escherichia coli* K12 MG1655 ETS variants used in this study are from the previous project (35,72). Three independent lineages of each ETS variant were evolved on two separate M9 minimal media preparations, one with glycerol (0.2%) and another with succinate (2 g/l) as the carbon source. Adaptive laboratory evolution was performed as described in previous work (35).

3.7.2 DNA sequencing and RNA sequencing

A clone from the evolved strains was used for DNA and RNA sequencing. Total DNA was extracted from an overnight grown culture, and total RNA was extracted from a culture at an A600 ~0.6. DNA isolation, library preparation, and subsequent analysis were performed as previously described (114). RNA isolation was performed using the Qiagen RNeasy Mini kit (Cat. No. 74104), following the manufacturer's protocol of "enzymatic lysis of bacteria." Ribosomal RNA was removed using the NEBNext rRNA depletion kit Bacteria (Cat. no. NEB E7850X), and cDNA libraries were constructed using the NEBNext Ultra II Directional RNA Library prep kit for Illumina (Cat. no. E7760L).

3.7.3 Phenotype characterization

Exometabolites were estimated from the culture media. Samples were filtered through a 0.22 μm filter (PVDF, Millipore) and measured using refractive index detection by HPLC (Agilent 1260 Infinity) with a Bio-Rad Aminex HPX87-H ion exclusion column. The HPLC method was the

following: injection volume of 10 µL and 5 mM H_2SO_4 mobile phase set to a 0.5 mL/min flow rate at 45°C. The phenotype dataset was used for the aero-type classification of the strains as described previously (72).

3.7.4 Omics-constrained metabolism and macromolecular expression modeling

The ATS proteome allocation calculation employed here follows the same protocol described in Anand *et al.* 2022 (35). Full details of the procedure, including ATS proteome allocation calculations, ATP generation reaction fluxes, and metabolic flux mapping, can be found in the cited publication. Minor modifications include using succinate or glycerol uptake rate as a constraint rather than glucose uptake rate. Additionally, the following BiGG metabolic reactions were used for identifying the quinone ETC fluxes (144): NADH16pp, NADH5, CYTBO3_4pp, CYTBDpp, NADH17pp, NADH10, CYTBD2pp, FDH4pp, and FDH5pp.

3.7.5 iModulon correlation analysis

iModulon analysis was performed using the iModulon framework (12,14). This platform applies Independent Component Analysis (ICA) to decompose RNA-seq data into iModulons, which represent independently regulated gene sets. Activity scores for 201 curated iModulons were calculated across 1,035 transcriptomic samples and are available in our PRECISE-1K compendium (12,14).

Correlation analysis was conducted on the iModulon activities to identify coordinated independent transcriptional modules. Pairwise Spearman correlations were computed to generate a correlation matrix, which was then hierarchically clustered to identify groups of iModulons with significant co-regulation. Varying thresholding distances of the matrix defined robust clusters, including the seven iModulons that formed the aerobicity stimulon. These clusters were cross-referenced with experimental metadata to assess their relevance to aerobicities. For each iModulon pair within this cluster, activity scores across all samples in the PRECISE-1K compendium were

scatter-plotted, with ETS variants grown on the three substrates and anaerobic samples overlaid for comparison. The anaerobic samples are: p1k_00175, p1k_00176, p1k_00073, and p1k_00074.

3.7.6 iModulon constrained ME modeling and analysis

ME modeling was used to investigate the impact of increasing fluxes through ArcA iModulon-associated genes on proteome allocation. Simulations were initiated using wild-type (WT) flux values derived from a strain growing with glucose as the carbon source. The fluxes associated with ArcA iModulon genes were progressively increased in increments of 2.5% relative to the WT flux, up to a maximum of 10% of WT flux, in rich media.

For each simulation step (WT, 1.025 * WT, 1.05 * WT, 1.075 * WT, and 1.010 * WT), the ME model calculated the proteome fraction allocated to the seven iModulons comprising the aerobicity stimulon. Total proteome allocation for each iModulon was calculated as follows:

$$Total\ Proteome\ Allocation = \sum_i mw_i * V_i^{translation}$$

Where mw_i and $V_i^{translation}$ represents the molecular weight and translation flux of the *i*th protein in the model.

3.7.7 Data and code availability

New transcriptomics accession codes for DNAseq SRA accession number: PRJNA1235627; RNAseq GEO accession number: GSE291717. Old transcriptomics accession codes: GEO accession number GSE202144. The full PRECISE1k decomposition matrices are available at https://github.com/SBRG/precise1k. Raw RNA-seq data used in PRECISE1k have been deposited at GEO and are publicly available. Accession numbers are listed in the metadata file located in the GitHub repository at the path: data/precise1k/metadata_qc.csv.

All the simulations performed in this manuscript can be reproduced using the FoldME model, which is constructed using the COBRApy toolbox version 0.5.11 for constraint-based modeling and its extension for ME-models, COBRAme version 0.0.9, ECOLIme version 0.0.9, and

solveME, all publicly available on Github (https://github.com/SBRG/cobrame, https://github.com/SBRG/ecolime, https://github.com/SBRG/solvemepy) (6,145). Custom code for constraining and solving ME-models can be found at https://github.com/SBRG/ME-script. Code for our Independent Component Analysis pipeline can be found on GitHub (https://github.com/SBRG/iModulonMiner, https://github.com/SBRG/precise1k), as well as code for all iModulon analysis (https://github.com/SBRG/pymodulon) (146).

3.8 Acknowledgments

This work was funded by the DAE, India-Tata Institute of Fundamental Research Grant to A.A. and the Novo Nordisk Foundation Grant Numbers NNF10CC1016517 and NNF20CC0035580 to B.O.P. We thank Marc Abrams (Systems Biology Research Group, University of California San Diego) for assistance with paper editing. Icons for some figures were created using our license at Biorender.com.

A.A., and B.O.P. designed the study. N.B., R.M., D.M.P., and A.A. performed experiments. A.P., N.B., S.K., A.M.F., B.O.P., and A.A. analyzed the data. A.P., and A.A. wrote the manuscript with contributions from all other co-authors.

This section of Chapter 3, in part, is a reprint of the material submitted to *bioRxiv*:

Patel A, Banwani N, Mink R, Prabhakaran DM, Khairnar SV, Feist AM, et al. Aerobicity stimulon in Escherichia coli revealed using multi-scale computational systems biology of adapted respiratory variants. bioRxiv; 2025. p. 2025.03.13.642450. Available from: https://www.biorxiv.org/content/10.1101/2025.03.13.642450v1

The dissertation author was the primary author.

Chapter 4. Integration of multi-omic knowledge-enriched data and computational tools across diverse case studies

The previous chapters established how knowledge-enriched data analytics and mechanistic modeling can each reveal distinct dimensions of cellular physiology: omics analyses capture the emergent structure of regulation and proteome allocation, while genome-scale ME models translate those insights into quantitative predictions of growth, fluxes, and protein costs. This chapter synthesizes a series of collaborative studies that demonstrate how these approaches can be combined across diverse biological contexts. Each study integrates experimental data, omics measurements, or independently derived knowledge constraints with mechanistic modeling to generate systems level insights that neither approach alone could achieve.

Together, these case studies illustrate a general principle of interoperability; the ability of ME models to exchange information with external data and analytics frameworks in a structured, quantitative manner. They encompass physiological trade-offs, environmental adaptations, and multi-scale regulatory responses spanning transcriptomics, proteomics, and metabolism. In each instance, ME modeling served as a quantitative backbone to interpret large datasets, evaluate proteome allocation, and assess how regulatory or environmental perturbations reshape cellular resource distribution.

These collaborative efforts also demonstrate how the interoperability framework developed throughout this dissertation generalizes to new organisms, stress conditions, and modeling paradigms. Across the following sections, mechanistic models were used to (i) reconstruct transcriptomic trade-offs between stress and growth, (ii) evaluate oxidative-stress tolerance mechanisms through integrative modeling, (iii) quantify proteome reallocation during high-density cultivation, (iv) assess energy-metabolism extensions such as extracellular electron transfer, (v) constrain ME-model predictions using omics-derived nutrient-response patterns and (vi) analyze

flux and proteome reallocation to strain with a knockout before and after evolution. Collectively, they show that combining omics-based knowledge enrichment with constraint-based mechanistic reasoning provides a coherent language for studying proteome allocation and stress physiology across scales.

4.1.1 The hallmarks of a trade-off in transcriptomes that balances stress and growth functions

The *E. coli* transcriptome exhibits a fundamental trade-off between stress preparedness and growth optimization. This study quantified that balance using iModulon analysis and tested its mechanistic basis through genome-scale metabolism and expression model simulations.

Maintaining optimal fitness in microorganisms requires navigating trade-offs in resource allocation (147). Expression profiling has shown that increasing ribosomal content ("greedy" genes) increases growth rate, while decreasing expression of stress-response ("fearful") genes is highly correlated with faster growth (147). Thus, the *fear-greed (f/g)* trade-off emerges as a key feature of transcriptome reallocation, in which cells that favor rapid growth face the cost of diminished responsiveness to stress (85).

The f/g trade-off was demonstrated as a property of the *E. coli* transcriptome through a data-driven approach defining independently modulated gene sets, termed iModulons (13). The trade-off is characterized by the strong negative correlation between the activity levels of the RpoS (fear) and Translation (greed) iModulons. This relationship involves competition between housekeeping and stress sigma factors (RpoD and RpoS), the stringent-response mediators ppGpp and DksA, and other regulatory mechanisms that act through RNA polymerase (RNAP) (148–150).

Further studies revealed that this transcriptomic allocation pattern is evolutionarily adjustable. In adaptive laboratory evolution (ALE) experiments, selection for rapid growth repeatedly yields RNAP mutations that shift expression toward greed (13,32,151–154). RNAP is one of the most frequent mutational targets during ALE (76), with variants arising across multiple

structural domains (155–163). The transcriptional impact of these mutations far from the catalytic core was unclear prior to this work. Here, a multi-scale analysis was used to examine how twelve distinct RNAP point mutations affect global expression and resource allocation.

RNAP mutations were introduced into the *E. coli K-12 MG1655* genome using CRISPR-based genome editing. RNA-seq analysis under aerobic growth on glucose M9 medium revealed that eleven of twelve mutations increased growth rate and caused a shift toward greed in the transcriptome. Structural modeling using PyRosetta (164) showed that many of these mutations destabilize the rpoB–rpoC interface, affecting the ppGpp-binding region (158).

All tested mutations led to down-regulation of stress-associated genes and up-regulation of translation-associated genes. iModulon analysis of the new transcriptomes confirmed that mutations in diverse structural regions produce similar regulatory outcomes: the RpoS iModulon activity decreased, while the Translation iModulon activity increased, reflecting the classic f/g trade-off. This convergence suggests that distinct molecular perturbations of RNAP channel the transcriptome toward a common resource-allocation state favoring growth over stress readiness (Figures 1-2 in Dalldorf *et al.*).

While transcriptomic analysis revealed relative shifts in iModulon activity, the magnitude of corresponding proteomic changes required quantitative modeling. A genome-scale ME model (94) was therefore deployed to reproduce the f/g trade-off and infer absolute proteome composition during this transition.

Simulations were run to maximize growth while constraining reactions associated with the RpoS iModulon to specified lower bounds. The resulting proteomic mass fractions for the RpoS and Translation iModulons were highly anti-correlated (-0.9994), indicating that as cells invest less in stress readiness, translational capacity expands nearly proportionally. A unit increase in Translation iModulon activity produced a ~650% stronger effect on its proteome fraction than on

the RpoS fraction, implying that small transcriptional shifts yield large phenotypic consequences. The model also predicted that forced expression of stress-response genes reduces growth-related protein synthesis, mirroring experimental observations (**Figure 4.1**).

These results show that the ME model framework quantitatively captures the proteome-allocation basis of the transcriptomic fear-greed axis. Integrating omics-derived iModulon constraints with mechanistic modeling thus bridges observed regulatory variation to its proteomic and energetic costs.

This multi-scale analysis establishes that small genetic perturbations in RNAP can reprogram transcriptome allocation along a conserved fear-greed axis. The combined use of iModulon analytics and proteome-constrained ME modeling provides a mechanistic foundation for interpreting such global trade-offs. The study further illustrates how expression-state decomposition can be interoperable with mechanistic modeling frameworks, forming a blueprint for the integrative approaches developed throughout this dissertation.

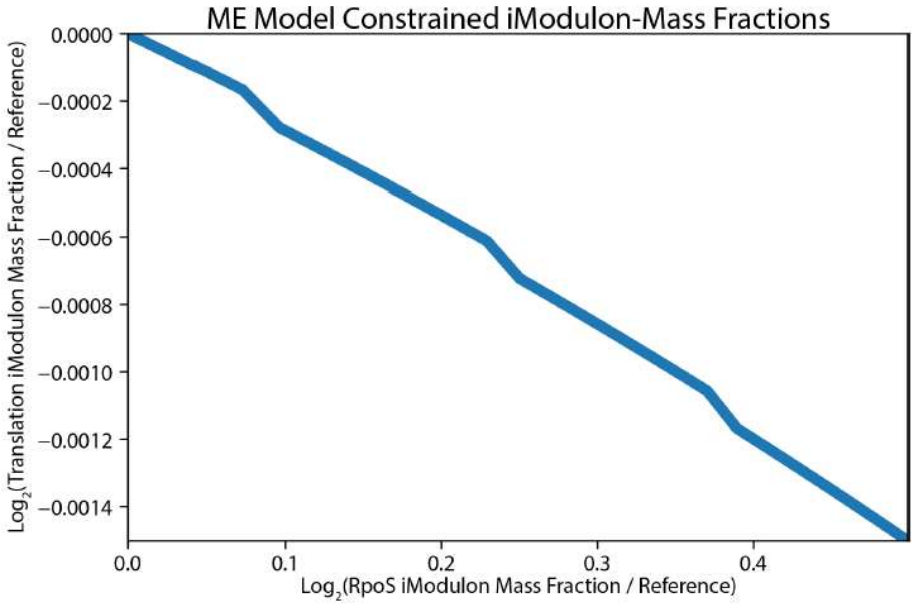


Figure 4.1: ME-model simulation of the fear-greed trade-off.

FoldME predictions showing anti-correlated proteome mass fractions of the RpoS (fear) and Translation (greed) iModulons under varying lower-bound constraints on stress-associated reactions (adapted from Dalldorf *et al.*, 2023, Supplemental Figure 2) (165).

4.1.2 Methods

4.1.2.1 iModulon Computation

RNA-sequencing data was used to create iModulon activity levels of the mutated strains using PyModulon which is available at https://github.com/SBRG/pymodulon (13). Activities of iModulons were compared to samples from PRECISE 2.0 (14) which is easily accessible using iModulonDB (24)https://pmc.ncbi.nlm.nih.gov/articles/PMC10120744/ - R33.

4.1.2.2 Metabolic Model and Proteomic Calculations

The FoldME (166) model was used for the metabolic modeling calculations. Figure 4.1 was generated by iteratively increasing the lower bounds for the genes of the RpoS iModulon and recording the proteomic mass fraction of the Translation and RpoS iModulons' genes until the model no longer ran. Proteome mass fraction to iModulon genes is the sum of the measured proteomic mass fractions of each enriched gene in an iModulon. This value is calculated for every sample and plotted against its corresponding PRECISE iModulon Activity. The proteomic calculations performed for this paper are well described in Patel et al (132)https://pmc.ncbi.nlm.nih.gov/articles/PMC10120744/ - R49.

4.1.3 Acknowledgments

This work was supported by The Novo Nordisk Foundation (NNF) Center for Biosustainability (CfB) at the Technical University of Denmark (NNF20CC0035580), National Institutes of Health Grant R01 GM057089, and the Y.C. Fung Endowed Chair in Bioengineering at the University of California San Diego. We would like to thank Marc Abrams (Systems Biology Research Group, University of California San Diego) for assistance with paper editing and José Utrilla (Center for Genomic Sciences, UNAM) for help with historical context.

C.D., and B.O.P. designed the study. R.S., and Y.H. performed experiments. K.R., B.O.P., C.D, and A.P. analyzed the data. C.D., and B.O.P. wrote the manuscript with contributions from all other co-authors.

This section of Chapter 4, in part, is a reprint of the material published in *mSystems*: Dalldorf, Christopher; Rychel, Kevin; Szubin, Richard; Hefner, Ying; **Patel, Arjun**; Zielinski, Daniel C; Palsson, Bernhard O (2024). *The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions.* mSystems, 9(7), e00305-24. The dissertation author was the fifth author of this material.

4.2.1 Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance

Oxidative stress imposes one of the most pervasive selective pressures in aerobic organisms. Paraquat, a redox-cycling compound that generates superoxide radicals, provides a controlled framework for studying how *Escherichia coli* reorganizes its transcriptome, proteome, and metabolism to mitigate reactive oxygen species (ROS) damage (167–169). In this study, multi-scale laboratory evolution and modeling approaches were combined to identify molecular mechanisms underlying paraquat tolerance and to quantify their proteomic and metabolic consequences.

Independent adaptive-laboratory-evolution (ALE) lines were propagated in M9-glucose minimal medium containing paraquat to induce chronic oxidative stress (170). End-point populations exhibited increased growth rate relative to their ancestors and convergent mutations in genes associated with redox homeostasis, transcriptional regulation, and central carbon metabolism.

RNA-sequencing of evolved clones revealed reproducible global expression changes captured through iModulon analysis. Activities of SoxRS, OxyR, and Fur iModulons increased sharply, reflecting up-regulation of superoxide dismutase, catalase, and iron-sequestration systems. In contrast, translation-related iModulons decreased in activity, consistent with a resource shift

from growth to protection. The modular decomposition highlighted a coordinated reallocation of transcriptional programs forming a quantitative description of oxidative-stress physiology.

To move beyond correlative transcriptomic patterns, a proteome-resolved ME-model framework (OxidizeME) (171) was applied to compute the quantitative cost of reactive-oxygen-species (ROS) damage and repair. The model explicitly represents iron–sulfur clusters, metalloproteins, and oxidative-damage pathways, allowing simulation of protein inactivation and repair fluxes under varying stress intensities.

Simulations constrained by experimental proteomic data reproduced the observed proteome reallocation between functional and damaged sectors (**Figure 4.2**). The model captured the redistribution of protein investment from biosynthetic and translational machinery toward detoxification, repair, and maintenance functions as paraquat stress increased. These predictions agreed with measured enzyme abundances and growth recovery following paraquat exposure, confirming that oxidative damage can be mechanistically encoded as an explicit constraint in genome-scale ME models.

The integration of *in vivo* omics measurements with *in silico* oxidative-damage accounting established a direct, quantitative mapping between stress-induced transcriptomic changes and proteome-allocation shifts. The same framework accurately predicted experimentally measured differences in enzyme abundances and growth recovery following paraquat exposure, confirming that ROS damage can be mechanistically encoded as an explicit constraint in genome-scale ME models.

Model-guided analysis revealed that oxidative tolerance arises from a composite of mechanisms: (i) reduction of intracellular redox cycling through down-regulation of electron-leak pathways; (ii) enhanced turnover of oxidized proteins; and (iii) proteome investment rebalancing from biosynthetic machinery to maintenance functions. This systems-level view reconciles the

trade-off between aerobic growth efficiency and stress robustness, an extension of the allocation principles introduced in the previous section.

Together, these results illustrate how coupling transcriptomic decomposition with ME-modeling frameworks such as OxidizeME enables mechanistic quantification of stress adaptation. The interoperable workflow (omics input, modular analysis, and proteome-constrained simulation) provides a blueprint for linking observed regulatory responses to their underlying biochemical and energetic costs.

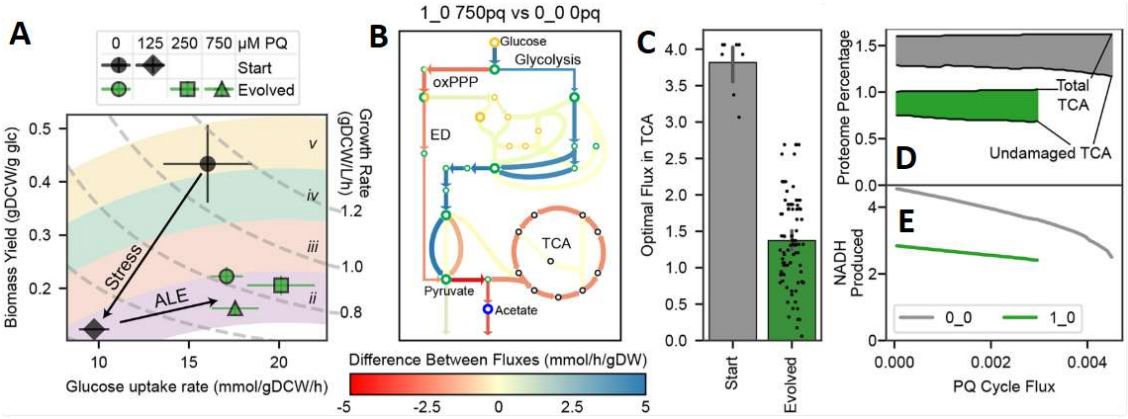


Figure 4.2: Proteome reallocation under oxidative stress.

Experimental and OxidizeME-predicted partitioning of functional and damaged proteome fractions in *E. coli* evolved for paraquat tolerance. A) Aero-type plot computed from experimental growth rates and glucose uptake rates. B) Flux differences comparing the starting strain and one evolved on high paraquat. C) Distribution of TCA cycle fluxes. D-E) Damage to TCA cycle enzymes and NADH production decreases under ROS stress (adapted from Rychel *et al.*, 2023, Fig. 6C–G) (172).

4.2.2 Methods

4.2.2.1 ME Modeling

We used OxidizeME, a genome-scale model of metabolism and expression (ME) with ROS damage responses (171). Models used for flux maps were constrained using phenotypic data (glucose uptake rate and growth rate) and expression data as previously described (34,35). In order to force PQ cycling in the model, the lower bounds for the 'PQ2RED_FWD_FLAVONADPREDUCT-MONOMER_mod_fad' and 'PQ1OX_FWD_SPONT'

were set to the same non-zero value and iterated over. Additionally, the former reaction was amended to accept NADH as an electron donor by editing the stoichiometry. PQ cycling sweeping calculations were performed by sampling various lower bounds to identify the range the model could support growth, and then sweeping 100 uniform values within that range. The total NADH produced through the TCA cycle was calculated by summing the fluxes for the 'MDH' and 'AKGDH' metabolic reactions. The percentage of the proteome allocated to the TCA cycle was calculated using the solutions from each model, specifically the translation fluxes:

$$\%\ Proteome\ Allocated\ to\ the\ TCA\ cycle = \frac{\sum_i mw_i * V_i^{translation}}{\sum_j mw_j * V_j^{translation}}$$

Where mw_i and $Vtranslationi$ represents the molecular weight and translation flux of the *i*th protein in the TCA cycle, and mw_j and $Vtranslationj$ represents the molecular weight and translation flux of the jth protein the entire model. The damaged portion of the proteome was calculated as follows:

$$\%\ Damaged\ Proteome\ Allocated\ to\ the\ TCA\ cycle = \frac{\sum_k mw_k * V_k^{comeplexformation}}{\sum_j mw_j * V_j^{translation}}$$

Where mw_j and $Vtranslationj$ are the same variables above,

and mw_k and $Vcomeplexformationk$ correspond to the kth protein in Supplemental Table S5. The undamaged portion of the proteome allocated to the TCA cycle was calculated as the difference between the total proteome allocated and the damaged proteome allocated.

4.2.3 Acknowledgments

This work was funded by the Novo Nordisk Foundation grant numbers NNF10CC1016517 and NNF20CC0035580, and by National Institute of General Medical Sciences grant number GM057089. It used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We would like to thank Marc Abrams for his

assistance with manuscript editing and Lachlan Munro for assistance with wet lab experiments. We would also like to thank Daniel Zielinski, Lei Yang, Emre Ozdemir, Tobias Alter, Ke Chen, and Jayanth Krishnan for helpful discussions. We thank Jason Yang for invaluable discussions and advice on an early draft of this manuscript. The graphical abstract was created in Biorender.

K.R., J.T, C.A.O., J.H.P, L.Y., A.M.F, and B.O.P. designed the study. J.T., J.H.P, R.S., Y.H., A.A., C.A.O., J.J., and E.T.T.M performed experiments. K.R., J.T., A.V.S., P.V.P., and C.L. analyzed the data. A.P. performed simulations. K.R., J.T., and B.O.P. wrote the manuscript, with contributions from all the other co-authors.

This section of Chapter 4, in part, is a reprint of the material published in *Cell Reports*: Rychel, Kevin; Tan, Justin; **Patel, Arjun**; Lamoureux, Cameron; Hefner, Ying; Szubin, Richard; Johnsen, Josefin; Mohamed, Elsayed Tharwat Tolba; Phaneuf, Patrick V; Anand, Amitesh; Olson, Connor A; Park, Joon Ho; Sastry, Anand V; Yang, Laurence; Feist, Adam M; Palsson, Bernhard O (2023). *Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance.* Cell Reports, 42(9).

The dissertation author was the third author of this material.

4.3.1 Trade-off between resistance and persistence in high cell-density cultures

High cell density growth imposes strong physiological constraints that force cells to balance resource investment between rapid proliferation and maintenance of stress resistance (173–177). This study combined quantitative experimentation with genome-scale modeling to define that trade-off and its proteomic basis in *Escherichia coli*.

Batch and fed-batch fermentations were conducted across increasing biomass densities to track changes in respiration, metabolite secretion, and stress-response gene expression. At elevated densities, oxygen limitation and by-product accumulation induced a gradual shift from high growth "resistant" states toward slow growth "persistent" phenotypes characterized by enhanced stress-

protective functions. Proteomic analysis confirmed that ribosomal and translational components declined, whereas chaperones, detoxification enzymes, and envelope-stabilizing proteins accumulated.

To quantify the proteomic costs underlying this transition, the OxidizeME framework was used to simulate proteome allocation under experimentally measured ROS scaling factors (**Figure 4.3**). The model explicitly accounted for protein repair and turnover reactions activated by oxidative stress. Simulated allocations reproduced the observed decline in biosynthetic capacity and the proportional expansion of stress-response sectors as cell density increased. This agreement indicated that maintenance functions consume a progressively larger fraction of the proteome, limiting translational throughput and explaining the experimentally observed slowdown of growth at high optical densities.

The modeling results demonstrated that persistence emerges not from a single regulatory switch but from proteome reallocation driven by energy and redox constraints. This mechanistic interpretation links the macroscopic resistance-persistence trade-off to molecular-level constraints on protein synthesis and repair.

Together, the experimental and computational analyses revealed a quantitative relationship between stress resistance and proteomic maintenance at high density. By integrating omics measurements with OxidizeME simulations, the study provided a mechanistic explanation for density-dependent persistence, illustrating again how interoperable multi-omic and ME model frameworks can resolve global resource-allocation phenomena.

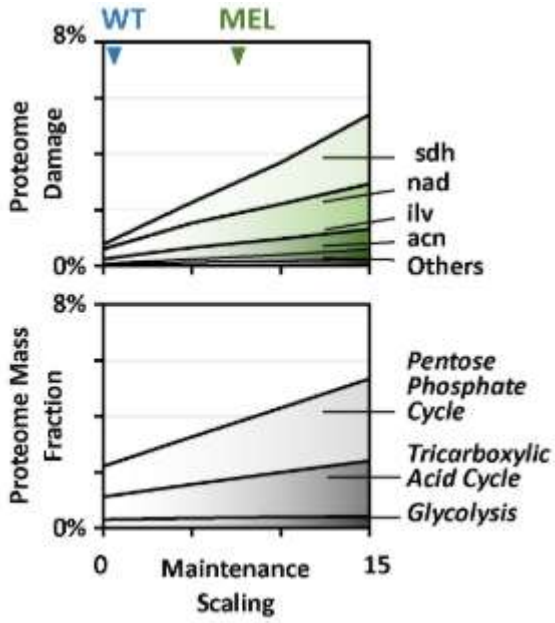


Figure 4.3: Proteome reallocation and damage in high-density cultures.
OxidizeME-predicted distribution of proteome sectors across ROS scaling factors as determined by experimental data (adapted from Beulig *et al.*, 2025, Fig. 3B) (178).

4.3.2 Methods

4.3.2.1 Genome-scale model of metabolism and macromolecular expression (ME model)

We used OxidizeME, a genome-scale model of metabolism and expression with ROS damage responses (171). Models were constrained using phenotypic data (growth rate, glucose uptake rate, oxygen uptake rate, CO_2 production rate, organic acid production rates) and expression data, as described previously (172). Growth was simulated over a range of maintenance costs (reaction "ATPM") with basal intracellular ROS concentrations for the WT strain (0.2 nmol L^{-1} superoxide and 50 nmol L^{-1} hydrogen peroxide). As the steady-state level of ROS is proportionate to the rate of its formation, intracellular ROS concentrations of the production strain were scaled according to the increase in oxygen uptake rate relative to the WT strain (169,171). The percentage of the proteome allocated to glycolysis, pentose phosphate cycle, and tricarboxylic acid cycle was calculated using the solutions from each model, according to the following:

$$\%\ \text{proteome} = \left(\sum_i \text{mw}_i \times V_{i,\text{translation}}\right) / \left(\sum_j \text{mw}_j \times V_{j,\text{translation}}\right)$$

where mw_j and $V_{j,\text{translation}}$ represent the molecular weight and translation flux of the *i*-th protein in a given pathway, respectively, and mw_j and $V_{j,\text{translation}}$ represent the molecular weight and translation flux of the *j*-th protein in the entire model. Similarly, the damaged portion of the proteome was calculated, according to:

$$\%\ \text{damaged proteome} = \left(\sum_k \text{mw}_k \times V_{k,\text{ComplexFormation}}\right) / \left(\sum_j \text{mw}_j \times V_{j,\text{translation}}\right)$$

where mw_k and $V_{k,\text{ComplexFormation}}$ represent the molecular weight and translation flux of all OxidizeME protein damaging reactions (denoted with the prefix "damage_"), respectively.

4.3.3 Acknowledgements

This work was supported by The Novo Nordisk Foundation Center for Biosustainability at the Technical University of Denmark under NFF grant numbers NNF20CC0035580 and NNF10CC1016517. We thank Søren Molin for valuable discussions and feedback on the manuscript.

F.B., J.B.-R., S.H.K., K.D., D.Z., L.Y., E.Ö., S.S. and B.P. designed the study. S.H.K. and L.Y. performed strain design and construction. F.B., J.B.-R., P.E.J., S.H.K., V.K., L.Y., E.Ö., and S.S. performed experimental design and implementation. J.B.-R., P.E.J., S.H.K., V.K., and C.S. performed experiments. K.D., D.Z., and E.Ö. performed initial data exploration. F.B. analyzed the data. F.B. and A.P. performed simulations. F.B. and B.P. wrote the manuscript with contributions from all the other co-authors.

This section of Chapter 4, in part, is a reprint of the material published in *mSystems*: Beulig, Felix; Bafna-Rührer, Jonas; Jensen, Peter E; Kim, SH; **Patel, Arjun**; Kandasamy, Vijayalakshmi; Steffen, CS; Decker, Katherine; Zielinski, DC; Yang, Laurence; Ozdemir,

Emre; Sudarsan, Suresh; Palsson, Bernhard O (2025). *Trade-off between resistance and persistence in high cell density Escherichia coli cultures.* mSystems, 10(7), e0032325.

The dissertation author was the fifth author of this material.

4.4.1 Extracellular respiration is a latent energy metabolism in E. coli

Respiration in *Escherichia coli* is typically confined to the cytoplasmic membrane, where inner-membrane dehydrogenases and terminal oxidases transfer electrons to soluble acceptors (179–181). In this study, a combination of genetic, electrochemical, and systems-level analyses revealed that *E. coli* possesses an unrecognized capability for extracellular electron transfer (EET), allowing cells to respire by passing electrons to external acceptors through native and heterologous nitroreductases (182–186).

This study demonstrated that expression of specific nitroreductases enabled measurable anodic currents under both aerobic and microaerobic conditions. Transcriptomic measurements confirmed broad remodeling of central-metabolic and respiratory pathways when cells engaged in EET. Key redox enzymes, including NADH dehydrogenase I, succinate dehydrogenase, and cytochrome bd oxidase, were down-regulated, while enzymes supporting flavin and quinone turnover were induced. These molecular changes established that EET acts as an alternative energy metabolism that supplements oxygen respiration by expanding the accessible redox range of the cell.

To quantify the energetic and proteomic consequences of EET, we applied a ME model of *E. coli* in which additional heterologous redox reactions were introduced to represent extracellular electron transfer. Simulations were performed under the same oxygen and electron-acceptor regimes used experimentally. The model predicted that enabling EET decreases flux through oxygen-dependent oxidases while partially relieving proteome burden associated with redox-balancing enzymes (**Figure 4.4**). This proteomic "relief" corresponds to a measurable reduction in

stress-related enzyme demand and improved allocation of translational capacity toward growth-associated sectors. Quantitatively, the model reproduced the experimentally observed increase in ATP yield and the shift in quinone-coupled dehydrogenase usage reported in the electrochemical assays.

These simulations demonstrated that extracellular respiration provides both energetic and proteomic advantages by redistributing cellular resources from endogenous detoxification and redox-maintenance functions to productive metabolism. Integrating omics data and mechanistic ME modeling thus supplied a coherent quantitative framework connecting molecular changes to the macroscopic phenotype of EET-enabled growth.

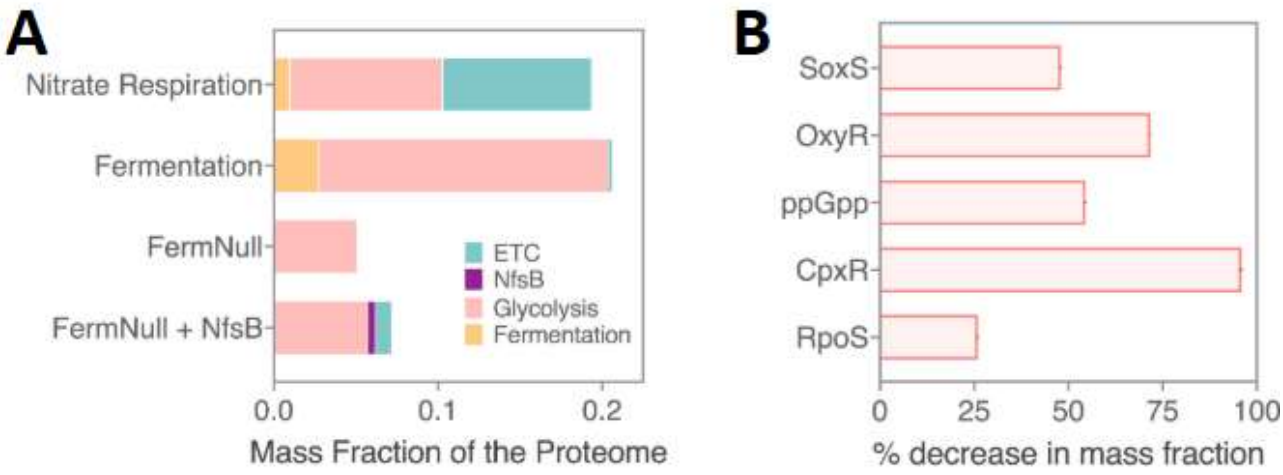


Figure 4.4: Proteome allocation and energetic relief during extracellular respiration.
ME-model predictions showing changes in proteome distribution between oxidative-stress and biosynthetic sectors when heterologous EET reactions are introduced, compared with experimental measurements (adapted from Kundu *et al.*, 2025, Fig. S6B–S6C) (187).

4.4.2 Methods

4.4.2.1 Modeling of HNQ-mediated EET

We performed Flux Balance Analysis using the latest genome-scale metabolic model of *E. coli*, iML1515 (188). We generated three sets of strains for analysis: 1) wild type, corresponding to the base iML1515 model; 2) EsinkNull and FermNull, which include the sets of redox

reaction knockouts described in this manuscript (see Strain list); and 3) EsinkNull (and FermNull) + nfsB, which has the EsinkNull (and FermNull) reaction sets knocked out but the addition of the reactions for naphthoquinone reduction by NADH (NADH + Oxidized HNQ ↔ NAD + Reduced HNQ), and regeneration by the anode (Reduced HNQ ↔ Oxidized HNQ). FBA was performed in a standard manner (189) using the COBRA Toolbox (190) in Matlab with the Gurobi solver version 9.1.1. We utilized growth as the primary objective for simulations, with glucose uptake constrained to 10 mmol/gDW/hr, oxygen constrained to 0 mmol/gDW/hr (anaerobic), with a secondary quadratic objective to minimize the Euclidean norm of the flux state while constrained to the maximal growth rate. Additionally, the NAD(P) transhydrogenase flux was capped to predicted wild-type levels of 2 mmol/gDW/hr in all simulations to prevent the use of futile redox cycles to oxidize NADH freely. Genes were knocked out respective of logical relationships in reaction gene-to-protein relationships (GPRs), such that a reaction flux was constrained to 0 mmol/gDW/hr if and only if all of its possible catalyzing protein complexes have at least one essential gene knocked out. Simulations were performed in Matlab using the COBRA toolbox (190).

We then analyzed the proteome allocation using the iJL1678b Metabolism and Macromolecular Expression (ME) model (6). We generated three strains for analysis: 1) wild type, identical to the base iJL1678b model; 2) FermNull, as described above; and 3) FermNull + nfsB, as described above. We estimated the NfsB k_{eff} value to be 57.4 in the model using the transcriptome-constrained wild-type flux state and experimental absolute proteomics from this study (132). All strains were grown anaerobically, with growth being the primary objective for simulations. The Wild Type strain was simulated under two conditions: 1) Fermentation, which was grown anaerobically with glucose uptake constrained to 10 mmol/gDW/hr, and 2) Nitrate Respiration, which was grown similarly but additionally with nitrate uptake constrained to 30 mmol/gDW/hr. The FermNull and FermNull + nfsB strains were constrained to the same growth rates calculated

from the iML1515 simulations. Proteome allocation was analyzed using the simulation results for ETC, Glycolysis, and Fermentation proteins. Proteome allocation related to stress iModulons (SoxS, OxyR, ppGpp, CpxR, and RpoS) were analyzed for the FermNull strains by calculating the proteome allocated to the genes enriched in each iModulons. Simulations were performed using the COBRAme Python software package (6).

4.4.3 Acknowledgments

We thank Dr. Rob Egbert from the Pacific Northwest National Laboratory for providing the BioDesignER strain. We thank Dr. Shyam Bhakta from Rice University for sharing the part vector used for constructing NfsB- and NfsA-expressing plasmids. We highly appreciate Dr. Steffen Lindner-Mehlich from the Max Planck Institute of Molecular Plant Physiology for valuable discussions on generating the NADPH auxotroph strain. We are also deeply grateful to Dr. Jayashree Soman from Rice University for proofreading the manuscript. The study was funded by Office of Naval Research, award # N00014-24-1-2034 (C.M.A.-F.); Cancer Prevention and Research in Texas, award # RR1900063 (C.M.A.-F.); and Novo Nordisk Foundation, award # NNF20CC0035580 (B.O.P.).

Project conceptualization, B.B.K. and C.M.A.-F.; methodology, B.B.K., J.K., D.C.Z., and C.M.A.-F.; investigation, B.B.K., R.S., J.K., A.P., and D.C.Z.; visualization, B.B.K.; funding acquisition, C.M.A.-F. and B.O.P.; supervision, C.M.A.-F., D.C.Z., and B.O.P.; writing – original draft, B.B.K., J.K., R.S., A.P., D.C.Z., and C.M.A.-F.; writing – review & editing, B.B.K., D.C.Z., and C.M.A.-F., with inputs from everyone.

This section of Chapter 4, in part, is a reprint of the material published in *Cell*:

Kundu, Biki Bapi; Krishnan, Jayanth; Szubin, Richard; **Patel, Arjun**; Palsson, Bernhard O; Zielinski, Daniel C; Ajo-Franklin, Caroline M (2025). *Extracellular respiration is a latent energy metabolism in Escherichia coli.* Cell, 188(11), 2907–2924.e23.

The dissertation author was the fourth author of this material.

4.5.1 Escherichia coli chooses substrates using a multi-layered hierarchical regulatory network

Carbon utilization in *E. coli* follows a structured regulatory cascade rather than a binary catabolite-repression switch (91,191–195). This study compiled 881 transcriptomes across 43 single-carbon substrates, identifying a four layer hierarchy that links substrate quality, growth rate, and proteome demand (196–198). Independent component analysis resolved 137 iModulons, of which 25 governed carbon catabolism, grouped into six clusters (I-VI) representing progressive regulatory layers.

Preferred sugars (Group A) were metabolized through CRP-independent modules, while secondary substrates (Group B) required three modular CRP sub-programs. As carbon quality declined, NtrC-1 and Propionate iModulons (Cluster IV) engaged to detoxify propionyl-CoA via the methylcitrate cycle, mitigating TCA-cycle overflow stress (199–202). A final salvage layer, sgcABCEQX, was induced only when carbon starvation coincided with ammonium **excess**, indicating a latent phage-borne route for pentitol or pentose utilization (203–206). Diauxic and starvation time-courses confirmed the dynamic reversibility of this four-layer network (207).

To quantify the cost of traversing these layers, genome-scale metabolism and expression model simulations (6) were performed using omics-derived constraints from each substrate condition. These analyses reproduced the hierarchical transcriptomic pattern and measured the proteome demand for carbon utilization (**Figure 4.5**). The required proteome mass fraction increased ~9.4-fold from Group A to D, with CRP-linked programs accounting for ~5% and RpoS-mediated stress functions requiring an additional 5.5-10.5%, for a total proteome reallocation of up to ≈ 16% on poor substrates. This explicit quantification, which formed my contribution,

demonstrated how modular omics analysis can directly constrain mechanistic ME models to capture rate-yield and stress-allocation trade-offs.

By uniting transcriptomic decomposition with proteome resolved modeling, this study provides a quantitative, interoperable framework that links substrate chemistry to regulatory hierarchy to proteome economy. It exemplifies the central theme of this dissertation: that multi-omics knowledge and mechanistic models can be integrated to elucidate and predict how *E. coli* balances metabolic breadth, energetic yield, and stress preparedness across environments.

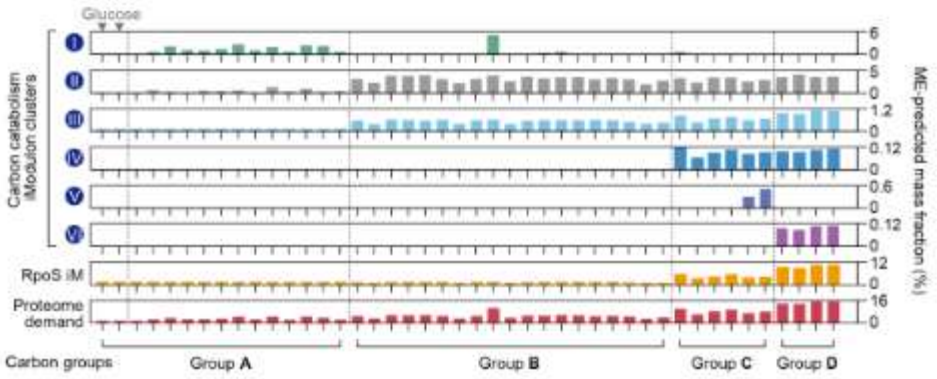


Figure 4.5: Proteome allocation across the nutrient-response hierarchy.
ME model predictions constrained by omics data reproduce the monotonic rise in proteome investment from Group A to D substrates (adapted from Shin *et al.*, 2025, Figure 5C).

4.5.2 Methods

4.5.2.1 Omics-constrained metabolism and expression (ME) model simulations

Proteome allocation followed the previous study (35) using the FoldME model (166) implemented in COBRAme (6). Experimental growth rates constrained the model; expression data were integrated as described previously (35), and steady-state solutions provided protein mass fractions assigned to carbon utilization-related and RpoS iModulons. Proteome sectors that were constrained using expression data include: Crp-1/2/3, NtrC-1, Propionate, DmlA, SCFA, SgcABCEQX, ppGpp, Pentose Phosphate Pathway, TCA Cycle, Oxidative Phosphorylation, and Glycolysis.

4.5.3 Acknowledgments

This work was supported by The Novo Nordisk Foundation (NNF) Center for Biosustainability (CfB) at the Technical University of Denmark (NNF20CC0035580) and the Y.C. Fung Endowed Chair in Bioengineering at the University of California San Diego. Additional support was provided by Global-Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2024-00442775).

Most wet-lab experiments were performed at the Department of Bioengineering, University of California San Diego. Subsequent validation experiments, advanced data analysis, and manuscript preparation were carried out after J.O.S. relocated to the Department of Biological Sciences, Chonnam National University. J.O.S. and B.O.P. conceived the study. J.O.S. and B.O.P. supervised the study. J.O.S. and B.O.P. designed the experiments. J.O.S., Y.H., R.S., and J.M.S. performed the experiments. J.O.S., A.P., X.A.L., E.C., J.K., H.F.S., D.Z., and B.O.P. analyzed the data. J.O.S., D.Z., and B.O.P. wrote the manuscript. All authors read and approved the final manuscript.

This section of Chapter 4, in part, is a reprint of the material submitted to *Nature Microbiology*:

Shin, Jongoh; **Patel, Arjun**; Lou Xuwen A; Catoiu, Edward; Krishnan, Jayanth; Hefner, Ying; Szubin, Richard; Sung, Jaemin; Francis Son, Hyeoncheol, Zielinski, Daniel C; Palsson, Bernhard O (2025). *Escherichia coli chooses substrates using a multi-layered hierarchical regulatory network*.

The dissertation author was the second author.

4.6.1 Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications

The pyruvate dehydrogenase complex (PDC) forms the primary link between glycolysis and the tricarboxylic-acid (TCA) cycle, converting pyruvate to acetyl-CoA (114,143,208). In *E. coli*, its transcriptional regulator PdhR represses PDC expression; deletion of *pdhR* results in uncontrolled overproduction of this 4.5 MDa complex and a pronounced growth defect under aerobic conditions (110,209). Genome-scale ME-model simulations (iJL1678b-ME) were used to quantify the systems-level cost of excess PDC synthesis. As PDC formation was varied from 0.1 × to 20 × its optimal rate, simulated growth declined monotonically, revealing that the proteome cost of assembling such a large multienzyme complex directly limits cellular fitness. When the model was re-parameterized to represent a minimal three-subunit "small PDC," the predicted growth penalty was substantially reduced, confirming that macromolecular size amplifies expression burden.

Adaptive laboratory evolution (ALE) of the Δ*pdhR* strain for ~300 generations restored growth to > 0.9 h^{-1}. Whole-genome sequencing showed convergent mutations in the Shine-Dalgarno sequence of *aceE*, the E1 subunit of PDC, decreasing ribosome-binding affinity and thereby re-balancing PDC expression (210). Transcriptomic and enzymatic assays confirmed reduced *aceE* expression and enzyme activity to near-wild-type levels. Thus, translation-level tuning compensated for the loss of transcriptional regulation.

Comparative transcriptomics revealed that evolved Δ*pdhR* replicates converged toward the expression profile of the glucose-optimized wild type. Re-established growth coincided with re-wiring of the electron-transport chain: type II NADH dehydrogenase returned to baseline, while cytochrome bd-II and pyruvate oxidase (PoxB) were up-regulated, expanding periplasmic electron-flow capacity and mitigating reactive-oxygen stress. Independent-component analysis showed elevated activity of the RpoS iModulon, with induction of its anti-adapter *iraP* and repression of *nadB*, collectively enhancing oxidative-stress tolerance without sacrificing growth. This pattern

reflected an alternate “fear-greed” trade-off in which Δ*pdhR* ALEs maintained RpoS-driven stress preparedness while regaining translational efficiency (211).

To mechanistically interpret these shifts, ME-model simulations using the FoldME framework (166) were performed by constraining the model with measured exchange fluxes, growth rates, and transcript-derived pathway mass fractions. The resulting flux maps (**Figure 4.6**) showed increased TCA-cycle throughput and oxidative efficiency in the evolved strains, consistent with higher ATP yield and the absence of overflow acetate secretion. Unlike wild-type ALEs that switch to the Entner-Doudoroff pathway, the evolved Δ*pdhR* lines retained Embden-Meyerhof-Parnas glycolysis, implying that RpoS-mediated redox resilience allowed continued use of the higher ATP yield route.

Together, these results demonstrate that dysregulation of a major proteome-intensive complex creates a quantifiable systems-level cost that can be resolved by translational fine-tuning and metabolic reallocation. The ME modeling analysis linked experimental transcriptomic and flux data to the predicted redistribution of proteome resources and energetic fluxes, establishing a mechanistic bridge between molecular regulation, redox balance, and growth optimization. This case study exemplifies the core theme of this dissertation: how interoperable integration of omics measurements and genome-scale ME models reveals the quantitative proteome-economics underlying adaptive evolution.

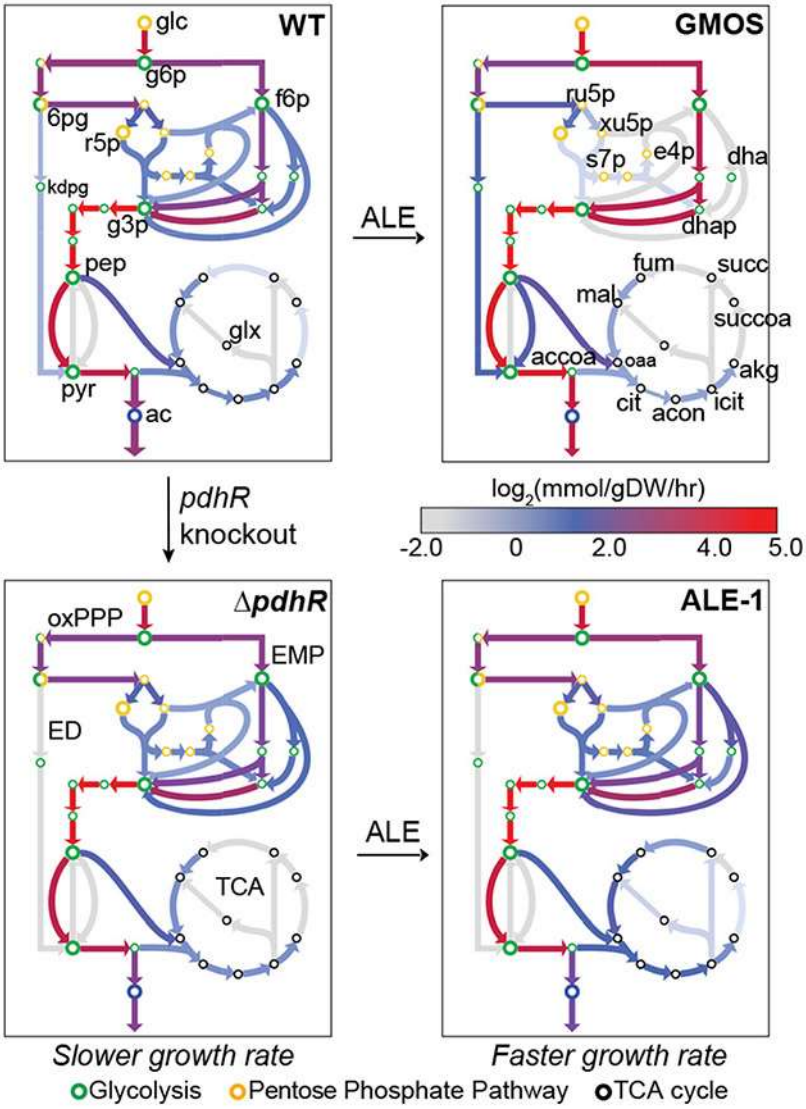


Figure 4.6. Optimization of glycolytic and TCA efficiency
Computed metabolic flux maps depicting the central metabolism in the WT GMOS, the pre-evolved, and the evolved Δ*pdhR* strains. Key metabolites indicated in the figure are as follows: glc, glucose; g6p, d-glucose-6-phosphate; g3p, glyceraldehyde-3-phosphate; f6p, d-fructose-6-phosphate; 6pgc, 6-phospho-d-gluconate; r5p, α-d-ribose 5-phosphate; pep, phosphoenolpyruvate; pyr, pyruvate; ac, acetate. Calculated fluxes of each strain are colored on a log scale (adapted from Anand *et al.*, 2021, Fig. 4) (34).

4.6.2 Methods

4.6.2.1 Metabolic Flux Mapping

We used FoldME, a genome-scale model of metabolism and protein expression, to elucidate the metabolic shifts between the WT, ΔpdhR, and evolved strains. FoldME was used for ME-model

simulations as it has shown to provide accurate depictions of proteome composition through the integration of the protein-folding network for E. coli (166).

The model was constrained with all phenotypic characterization data (i.e., growth rates and exchange rates), in addition to transcript mass fractions of major metabolic pathways calculated from RNA-seq data. Imposing the mass fractions directly onto the model yielded infeasible solutions due to disparities between the mass fractions and phenotypic data. Considering that the mass fraction of ribosomal proteins scales linearly with growth rate (212,213), we generated the following constraints for integrating the transcriptomic profile:

$$\phi_r \cdot V_{pathway} \geq \phi_{pathway} \cdot V_r$$

$$V_{pathway} = \sum_i mw_i \cdot V_i^{translation}$$

$$V_r = \sum_i mw_{rProtein_i} \cdot V_{rProtein_i}^{translation}$$

Where ϕ_r and $\phi_{pathway}$ represent the transcript mass fraction of ribosomal protein and the major pathway of interest, respectively. Additionally, mw_i and $Vtranslationi$ represent the molecular weight and translation flux of the *i*th protein in the corresponding pathway, whereas $mw_{rproteini}$ and $VtranslationrProteini$ represent the molecular weight and translation flux for the *i*th ribosomal protein.

4.6.3 Acknowledgments

This work was funded by the Novo Nordisk Foundation grant no. NNF10CC1016517 and National Institutes of Health grants R01GM057089 and U01AI124316. We would like to thank Marc Abrams (Systems Biology Research Group, University of California, San Diego) for his assistance with manuscript editing.

A.A. and B.O.P. designed the study. A.A., C.A.O., and R.S. performed the experiments. A.A., A.V.S, A.P., L.Y., and A.M.F. analyzed the data. A.A. and B.O.P. wrote the manuscript, with contributions from all the other co-authors.

This section of Chapter 4, in part, is a reprint of the material published in *Cell Reports*: Anand, Amitesh; Olson, Connor A; Sastry, Anand V; **Patel, Arjun**; Szubin, Richard; Yang, Laurence; Feist, Adam M; Palsson, Bernhard O (2021). *Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications.* Cell Reports, 35(1), 108961.

The dissertation author was the fourth author of this material.

Chapter 5. COBRAme.org: A web application for solving and constraining metabolism and expression models

Genome-scale models of metabolism and expression (ME models) extend traditional flux balance analysis by coupling transcription, translation, and metabolism within a single stoichiometric framework. These models enable quantitative prediction of proteome allocation, enzyme usage, and growth physiology, but their technical complexity has limited adoption. Running ME simulations typically requires local software installation, substantial computational resources, and familiarity with Python based modeling frameworks such as COBRAme.

We developed the COBRAme Web Application, an open, browser based platform for running genome-scale ME simulations without requiring programming or local installation. Users can upload or select pre-existing ME models, modify exchange bounds, kinetic parameters, and gene knockouts, and optionally apply transcriptomic or proteomic constraints. The system architecture consists of a lightweight frontend, a backend job queue, and a worker hosted on Amazon Web Services (AWS) EC2 that executes simulations and automatically emails results to users. Three modeling environments are supported: COBRAme, for steady-state simulations; StressME, for proteome allocation under environmental stress; and DynamicME, for time-course simulations of growth and metabolism. Example use cases demonstrate simulations of gene knockouts on alternative substrates, stress-dependent proteome reorganization, dynamic multi-carbon growth, and synthetic omics perturbations reproducing the fear-greed tradeoff.

The COBRAme Web App provides the first fully web-based environment for proteome-constrained ME modeling. By separating the user interface from the computational backend, the platform ensures reproducibility and scalability while eliminating software barriers. It enables students, researchers, and educators to explore ME-model behavior interactively and serves as a

foundation for future deployment of multi-omics-integrated modeling frameworks. The web application can be accessed at https://cobrame.org.

5.1 Background

Genome-scale models of metabolism have become central tools in systems biology for predicting cellular physiology, gene essentiality, and metabolic responses across environments (214–216). The most widely used framework, flux balance analysis (FBA), enables constraint-based simulation of steady-state flux distributions and has been instrumental in reconstructing and analyzing hundreds of metabolic networks (189,217). However, FBA models neglect the molecular processes underlying enzyme synthesis, limiting their ability to capture proteome allocation and expression costs.

Metabolism and Expression (ME) models extend traditional metabolic reconstructions by explicitly coupling transcription, translation, and metabolism within a unified stoichiometric framework (6,127). This coupling allows direct simulation of proteome-constrained cellular physiology, providing mechanistic predictions of growth rate, enzyme usage, and resource allocation under genetic and environmental perturbations (218). While ME models have become increasingly powerful and generalizable, they remain technically difficult to use. Running even a single simulation often requires specialized software installation, substantial memory and CPU resources, and familiarity with Python based frameworks such as COBRAme (6). These barriers restrict broader adoption of ME modeling beyond a small community of developers and computational experts.

To address this accessibility gap, we developed the **COBRAme Web Application**, an open and interactive platform for running genome-scale ME model simulations directly through a browser interface. The application enables users to upload or select pre-existing ME models, adjust parameters such as exchange bounds, kinetic constants, or gene knockouts, and apply optional

omics-based constraints. Each simulation is executed remotely by a worker system hosted on AWS EC2, and the results, including fluxes, mass fractions, and visualization files, are automatically returned to the user by email. This architecture separates the user interface from the computational workload, ensuring reproducibility and allowing the application to scale with available cloud resources.

The COBRAme Web App represents, to our knowledge, the first fully web-based environment for proteome-constrained modeling. By removing installation requirements and simplifying the execution pipeline, the platform lowers the barrier to entry for new users while maintaining compatibility with advanced ME frameworks such as StressME and DynamicME. In this way, the application serves both educational and research purposes: it provides an intuitive starting point for exploring proteome allocation and a practical interface for running complex ME based simulations that previously required local execution.

5.2 Implementation

The COBRAme Web Application was designed to make genome-scale ME modeling accessible to a broad range of users while maintaining a robust and reproducible computational workflow (**Figure 5.1**). The system architecture consists of three main components: a user facing frontend, a backend storage and job queue, and a worker process that executes model simulations.

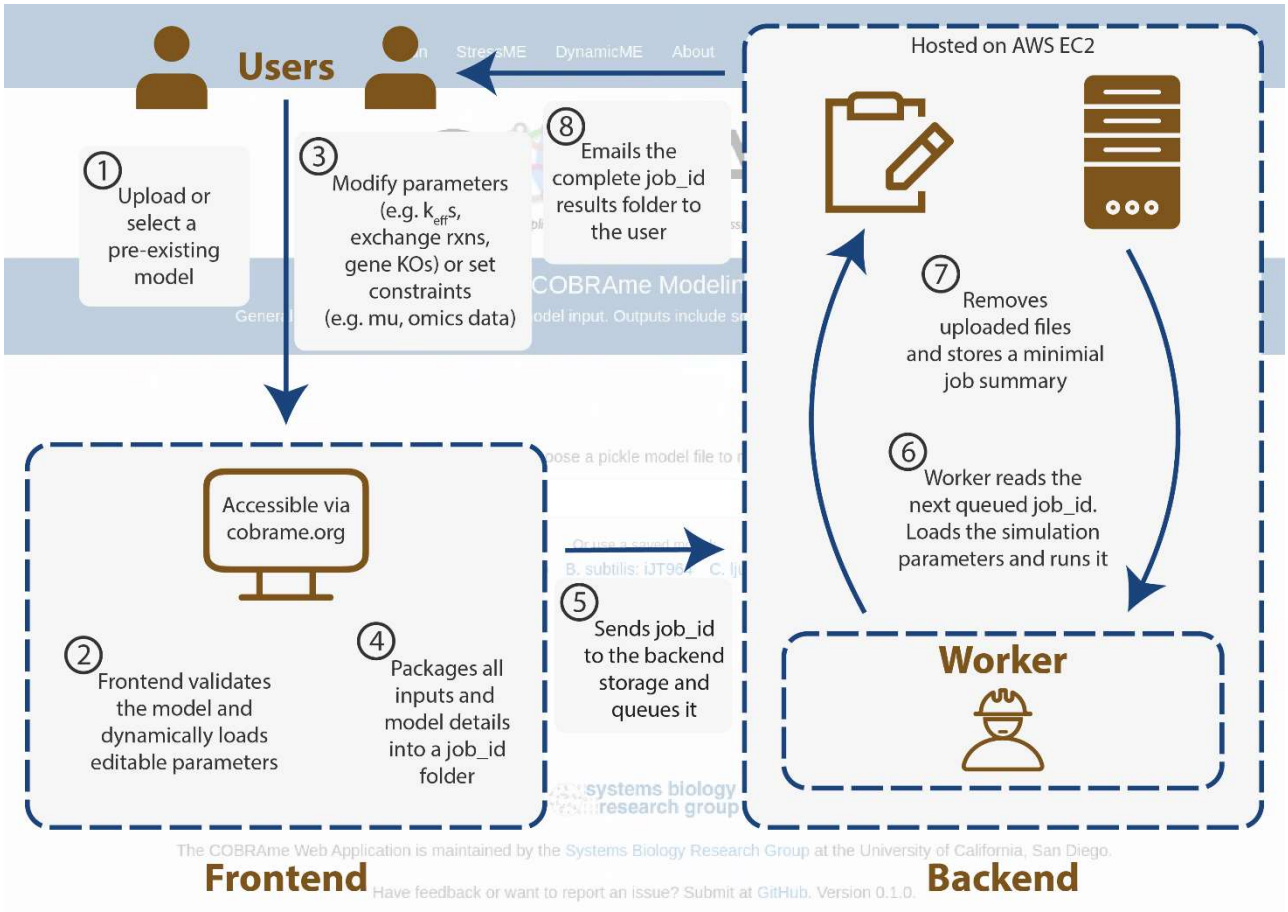


Figure 5.1: System architecture of the COBRAme Web Application.

The COBRAme Web App provides a complete end-to-end workflow for running proteome-constrained ME-model simulations directly through a browser interface. (1) Users upload or select a pre-existing ME model. (2) The frontend validates model structure and dynamically loads editable parameters. (3) Users modify kinetic parameters, exchange reactions, gene knockouts, or omics-based constraints (e.g., growth rate, transcriptomics, or proteomics). (4-5) The application packages all inputs into a job specific folder and sends it to the backend. (6-7) A worker instance hosted on AWS EC2 reads queued jobs, executes the ME simulation, stores minimal job summaries, and removes temporary files. (8) The full results folder is automatically emailed back to the user. This architecture separates user interaction from computationally intensive simulation to ensure reproducibility, scalability, and ease of use.

The frontend provides an intuitive browser interface for uploading, modifying, and submitting ME models. Users can either upload a saved model file or select from the three pre-implemented genome-scale ME reconstructions: *Escherichia coli* (iJL1678b-ME) (6), *Bacillus subtilis* (iJT964-ME) (219)https://paperpile.com/c/pG02AG/BMtv, and *Clostridium ljungdahlii* (iJL965-ME) (220). Once a model is loaded, the application dynamically retrieves editable parameters—including exchange reactions, gene knockouts, kinetic rate constants (k_{eff}), and

optional omics-based constraints such as growth rate or transcriptomics/proteomics data. All user inputs are validated on the client side to prevent incompatible model configurations prior to job submission.

When a job is submitted, the frontend packages all inputs and model metadata into a uniquely labeled job_id folder and transmits it to the backend. The backend manages a simple queue that records pending simulations and organizes associated job files. This design ensures that each simulation remains fully self-contained, improving reproducibility and facilitating downstream data retrieval.

A separate worker instance periodically scans the backend queue for the next available job. Upon retrieval, it loads the model, applies all specified parameters, and executes the simulation. The worker environment mirrors the standard COBRAme Python setup and can be deployed using the StressME Docker image for reproducibility and to maintain a consistent solver environment (221). Simulation results, including flux distributions, gene mass fractions, and summary statistics, are stored in a results subdirectory. Once complete, the worker automatically removes temporary input files, retains a minimal job summary, and emails the compressed results folder to the user.

The application currently supports three modeling environments that correspond to distinct ME modeling frameworks: COBRAme, StressME, and DynamicME (6,221,222). The general COBRAme interface performs steady-state simulations of ME models, allowing users to examine gene knockouts, substrate utilization, and proteome allocation under specified constraints. The StressME environment extends this workflow to include temperature, pH, and reactive oxygen species perturbations, integrating optional transcriptomic or proteomic data to compute stress induced proteome reorganization and generate flux maps and Proteomap visualizations. The DynamicME module enables time-course simulations, returning temporal trajectories of metabolite concentrations, biomass accumulation, and growth rate. Together, these three environments provide

a unified framework for exploring metabolic, stress, and dynamic dimensions of proteome-constrained modeling through a single web interface.

The entire application is hosted on an Amazon Web Services (AWS) EC2 instance, providing scalable computational resources and persistent storage. This separation of the user interface from the compute environment allows intensive ME simulations to run remotely without requiring local installation or software dependencies. Collectively, this architecture provides a stable, lightweight, and extensible foundation for running proteome-constrained models through a modern web interface.

5.3 Results and Disscussion

To demonstrate the functionality and versatility of the COBRAme Web App, we present several representative examples that can be executed directly through the browser interface. These examples highlight how users can configure and run ME model simulations, explore stress dependent proteome allocation, and perform dynamic time-course analyses without installing any software or writing code. The demonstrations are based on established ME model frameworks (COBRAme, StressME, and DynamicME) and are designed to illustrate the range of simulation types and outputs available through the platform.

5.3.1 General ME modeling workflow

The COBRAme environment provides an interactive interface for running steady-state ME model simulations on the ‘Main’ page, demonstrated here using the Escherichia coli iJL1678b model. In this example, the user simulates growth on fructose minimal medium while knocking out the *gnd* gene, which encodes 6-phosphogluconate dehydrogenase (**Figure 5.2a**). Within the web interface, the user locates *gnd*, or b2029, in the Gene Knockout box and selects it to set its state to “off.” A user can undo this by reclicking the blue chip that appears with the gene name. Exchange reactions are modified in the sortable Exchange Bounds table by setting the EX_fru_e lower bound

to –1000 mmol·gDW^{-1}·h^{-1} and the EX_glc__D_e lower bound to 0, effectively switching the carbon source from glucose to fructose.

The input tables are fully sortable and searchable, allowing users to quickly navigate hundreds of exchange reactions and confirm which metabolites are open or constrained. After defining parameters, the user enters an email address and submits the job. When complete, the backend sends a results archive containing all output files (**Figure 5.2b-c**). These include (**Figure 5.2d**): fluxes.csv, the complete set of metabolic reaction fluxes (mmol·gDW^{-1}·h^{-1}); mf.csv, the mass fraction of the proteome allocated to each gene product; and results_summary.json, the model result. The user also receives all input files, including a record of model modifications, ensuring the simulation can be precisely reproduced.

Generating a flux map shows decreased flux through the oxidative pentose phosphate pathway due to the *gnd* knockout and reduced growth yield on fructose relative to glucose. This example illustrates the reproducible, code free workflow for running conventional ME model simulations.

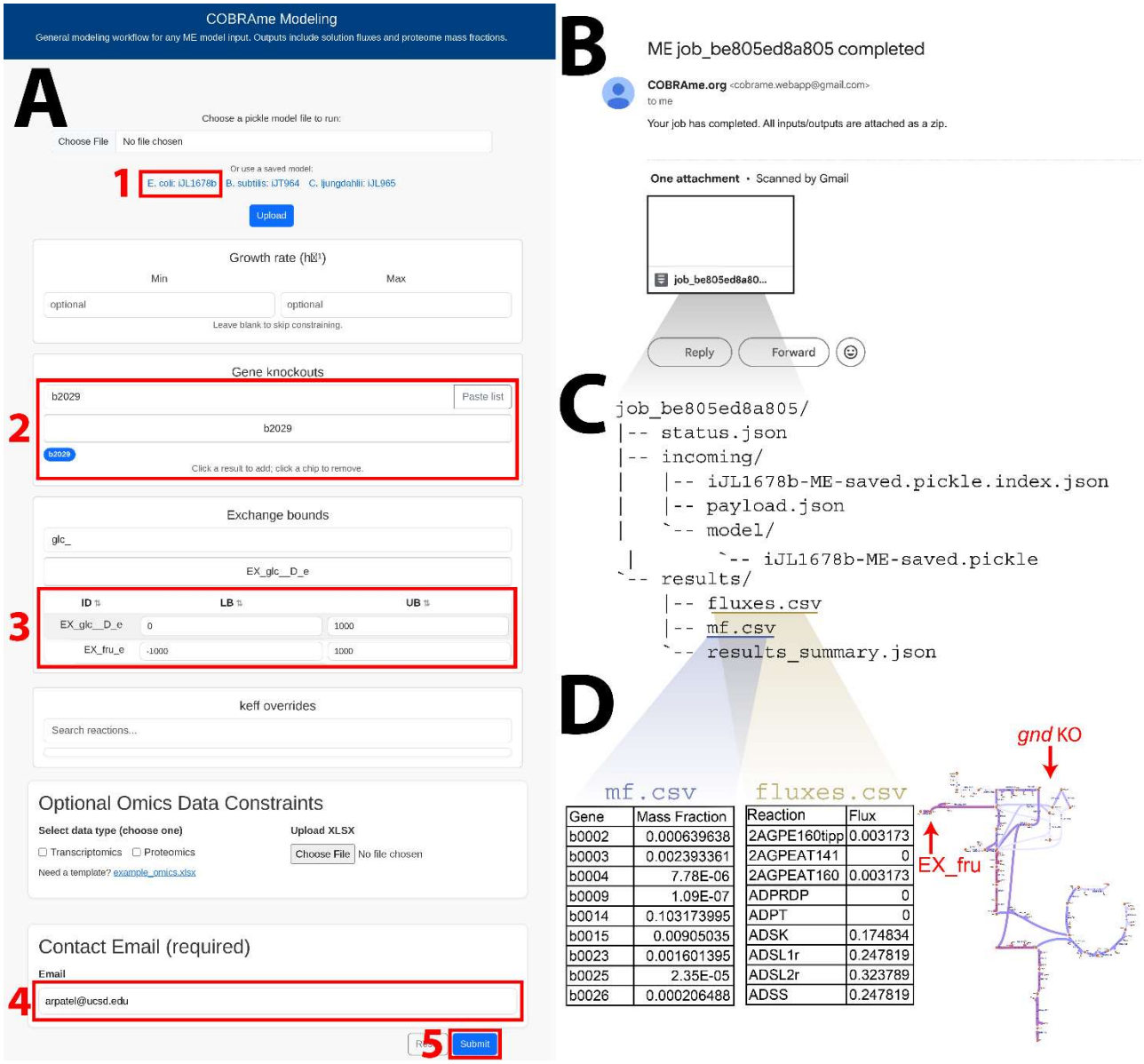


Figure 5.2: Interactive modeling workflow for the COBRAme module.

(A) Web interface for a simulation of E. coli iJL1678b with a gnd knockout and fructose as the carbon source. Users specify (1) the model, (2) gene knockouts, (3) exchange bounds, and (4) contact email before (5) submitting. **(B)** Automated email notification of job completion with a compressed results attachment. **(C)** File structure for a completed job, showing model, payload, and result directories. **(D)** Example output files, including mf.csv (mass fractions) and fluxes.csv (reaction fluxes), visualized as a network map highlighting the gnd knockout and fructose uptake reaction (EX_fru_e). Together these outputs demonstrate the app's ability to perform genome-scale ME simulations with user defined constraints.

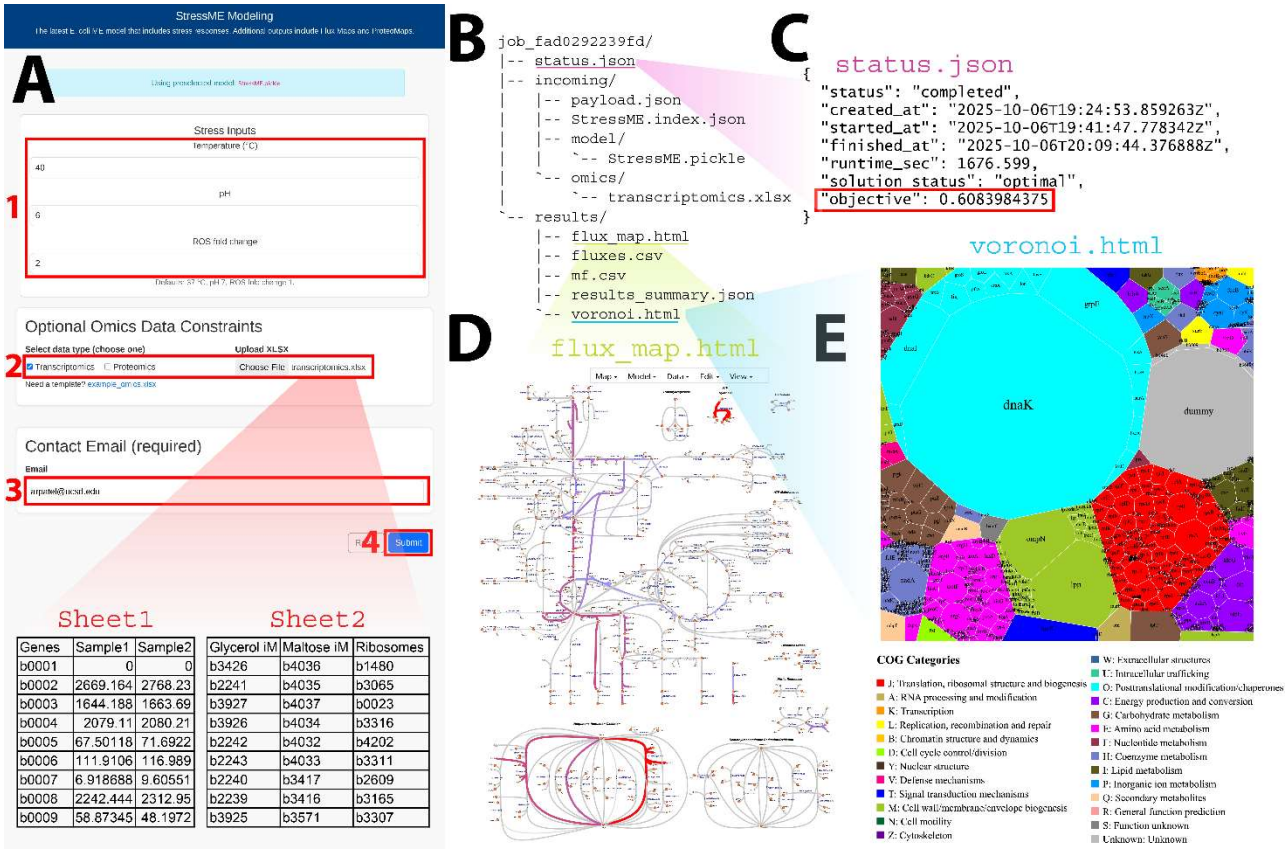
5.3.2 StressME modeling environment

The StressME module extends COBRAme to incorporate thermal, acid, and oxidative stress responses in a unified *E. coli* ME model (166,171,221,223). In this example, the user is on the 'StressME' page, which uses a pre-selected StressME model file. Three stress parameters are defined directly from dropdown menus: temperature = 40 °C, pH = 6, and ROS fold change = 2 (**Figure 5.3a**). The default values are set to wild type conditions.

In this example case, an optional transcriptomic dataset (transcriptomics.xlsx) is uploaded to constrain metabolic pathways. An example dataset is provided with clear instructions on how to format your data. The methodology for omic constraints uses a proteome sector approach to ensure the model produces macromolecules to at least the amount seen in the experimental data (15,16). We have used this method of omics integration in a variety of studies to showcase the benefits of layering additional experimental constraints (34,35,172,208,224).

The resulting job folder (**Figure 5.3b**) contains the core model file, omics constraints, a payload.json file that shows all user modifications/inputs, and a generated status.json that records model name, runtime, and final growth rate ($\mu = 0.608\ h^{-1}$) (**Figure 5.3c**). The flux_map.html output (**Figure 5.3d**) reveals metabolic flux distributions in an interactive escher map, where the user can change colors and thresholds. The associated voronoi.html file (**Figure 5.3e**) visualizes proteome allocation, showing a marked increase in stress response and chaperone sectors and a reduction in ribosomal mass fraction.

Together, these results demonstrate how StressME captures proteome redistribution under environmental challenges. By allowing users to toggle stress parameters and integrate omics data interactively, the module bridges experimental conditions with mechanistic predictions of proteome usage.



Genes	Sample1	Sample2
b0001	0	0
b0002	2669.164	2768.23
b0003	1644.188	1663.69
b0004	2079.11	2080.21
b0005	67.50118	71.6922
b0006	111.9106	116.989
b0007	6.918688	9.60551
b0008	2242.444	2312.95
b0009	58.87345	48.1972

Glycerol iM	Maltose iM	Ribosomes
b3426	b4036	b1480
b2241	b4035	b3065
b3927	b4037	b0023
b3926	b4034	b3316
b2242	b4032	b4202
b2243	b4033	b3311
b2240	b3417	b2609
b2239	b3416	b3165
b3925	b3571	b3307

Figure 5.3: StressME module for modeling proteome allocation under environmental stress**.**

(A) Frontend inputs for StressME simulations, including (1) temperature, pH, and ROS fold change, with (2) optional transcriptomic or proteomic data uploads, before the user enters their email (3) and submits (4). **(B)** Example job directory produced by the backend worker. **(C)** Representative status.json output containing runtime metadata and the final optimized objective value. **(D)** Generated flux_map.html, displaying flux distributions under stress conditions. **(E)** Corresponding voronoi.html Proteomap visualizing proteome allocation across COG functional categories. These outputs demonstrate integration of omics data and visual exploration of stress specific proteome allocation.

5.3.3 DynamicME modeling environment

The DynamicME environment performs time-course ME simulations, coupling transcription, translation, and metabolism to capture changes in cellular composition over time. Users define simulation parameters such as total duration = 10 h, time step = 0.1 h, reactor volume = 1 L, and initial biomass = 0.00675 gDW (**Figure 5.4a**). The initial concentrations table lists all extracellular metabolites with default values of 0 g/L. The table is sortable by metabolite name or value, allowing users to quickly identify which substrates are being included.

Users can optionally override individual catalytic rate constants (k_{eff}) in the Keff Overrides box. Each override can be reverted by clicking the red X next to the entry, restoring the default kinetic parameter from the model.

In this example, we wanted to recreate the multi-carbon substrate plot generated in the original DynamicME manuscript (222). In order to do this, we needed to include 0.4 g/L of gal_e, glc_e, malt_e, lac__L_e, glyc_e. We also want the output plots to track ac_e so we include a trace amount of 0.0001 g/L. The output plots only track metabolites that the user changes a value for. In order to replicate the DynamicME plot, we also need to modify the k_{eff}s of a handful of reactions (see **Supplemental Table D.1**).

Upon execution, DynamicME computes the full-time dependent trajectory of growth, substrate consumption, and byproduct secretion. The output directory (**Figure 5.4b**) includes the dynamicme_results.json file which includes a dictionary of all fluxes, biomass, concentrations, and mass fractions across all time points. Example plots (**Figure 5.4c-d**) plot biomass and growth rate over the simulation, as well as tracked metabolic concentrations over time. The interface allows direct inspection of temporal changes in proteome and metabolism, providing an intuitive framework for studying dynamic resource allocation.

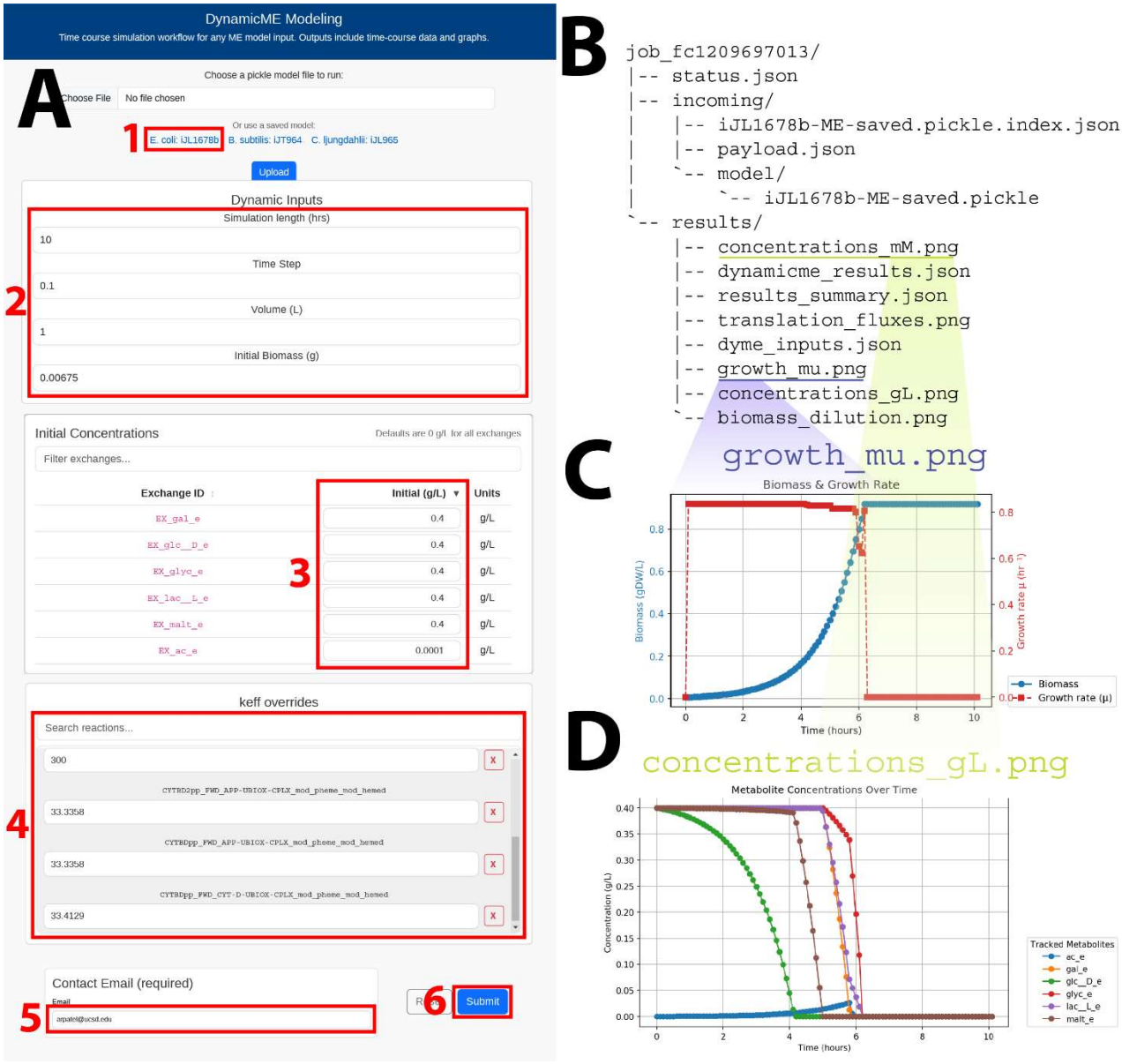


Figure 5.4: DynamicME module for time-course simulation of metabolism and expression.

(A) Frontend inputs specifying (1) the model, (2) dynamic simulation parameters (time step, volume, initial biomass), (3) initial metabolite concentrations, (4) keff modifications, (5) user email, and (6) submission. **(B)** Job output folder containing flux and concentration time-course results. **(C)** Example plot of biomass accumulation and growth rate over time. **(D)** Concentration trajectories for selected extracellular metabolites, illustrating substrate depletion and byproduct secretion during dynamic simulation. DynamicME enables temporal analysis of ME model behavior directly from the browser without requiring code execution.

5.3.4 Omics-constrained proteome tradeoffs

We applied the COBRAme Web App to reproduce the RpoS dependent “fear-greed” proteome tradeoff (165). In this example, the uploaded proteomic datasets were not derived from

experiments but rather constructed to deliberately shift proteome investment toward stress response genes. Multiple datasets were uploaded to represent incremental increases in RpoS gene group expression, 0%, 1%, 10%, and 20% of total protein mass, forcing the model to commit progressively more resources to RpoS associated functions (**Figure 5.5a**). The input spreadsheet (**Figure 5.5b**) lists the RpoS iModulon associated genes and their relative abundances under each condition.

After applying these synthetic constraints, simulations were executed sequentially, and the resulting mf.csv outputs were plotted to reveal proteome reallocation (**Figure 5.5c**). As the fraction of RpoS associated proteins increased, the mass fraction devoted to translational machinery decreased. Despite the artificial construction of these datasets, the resulting emergent tradeoff captures the same balance between growth efficiency and stress tolerance observed in experimental studies (165). This example illustrates how the COBRAme Web App can be used not only to analyze real omics data but also to design and test conceptual scenarios that probe resource allocation principles.

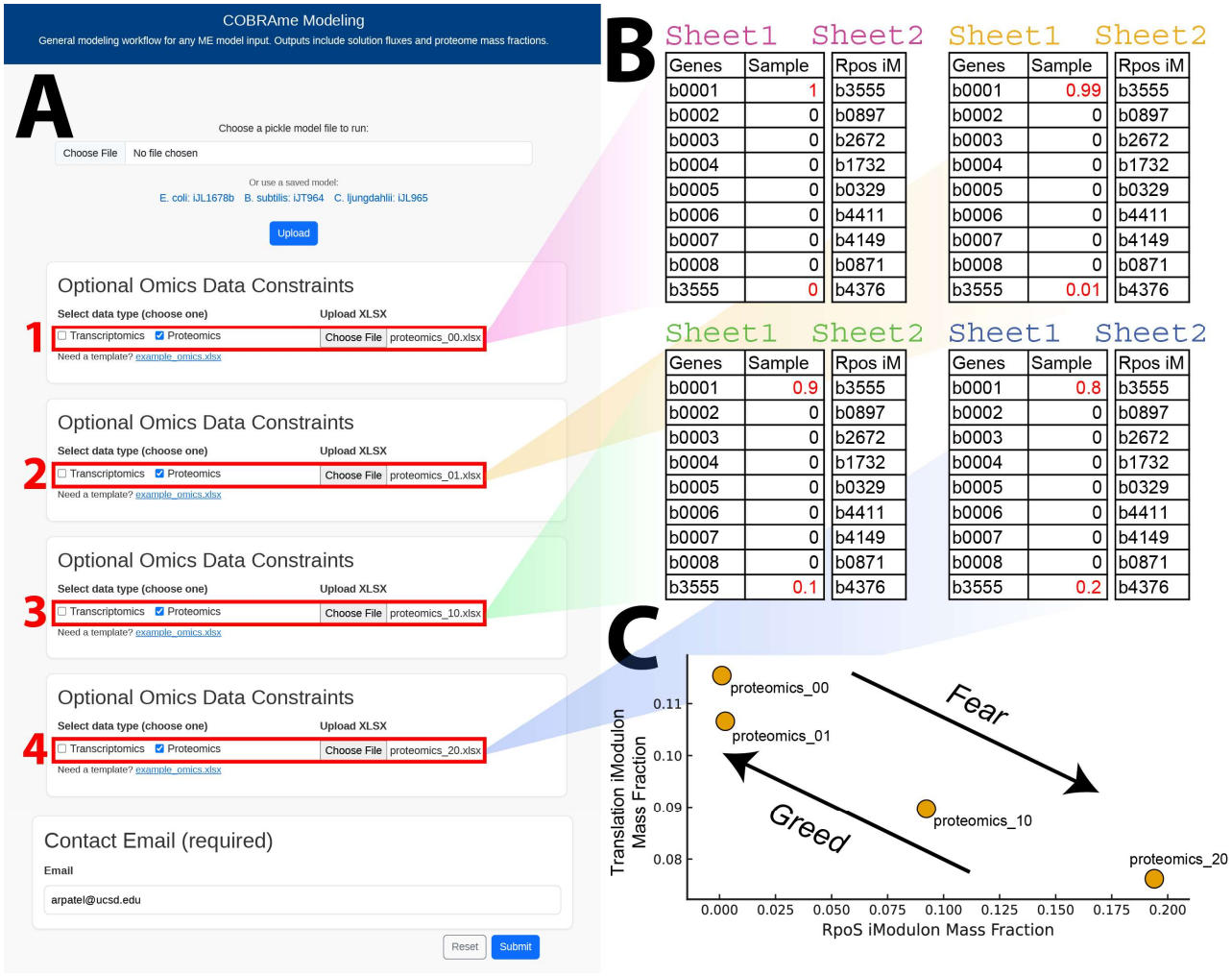


Figure 5.5: Reproducing proteome trade-offs using omics-constrained ME modeling.

(A) Web interface showing multiple proteomic dataset uploads for E. coli simulations under RpoS induced stress conditions. **(B)** Example spreadsheet inputs defining relative abundances of RpoS and translation related iModulons across conditions. **(C)** Modeled proteome mass fractions revealing the emergent "fear-greed" trade-off between RpoS and translation modules. This case study highlights how COBRAme Web App can reproduce experimentally observed proteome allocation principles through straightforward omics integration.

5.4 Conclusions

The COBRAme Web App enables interactive execution of genome-scale ME model simulations through a modern web interface. It allows users to upload, modify, and run models without requiring local software installation or programming experience. By integrating established frameworks such as COBRAme, StressME, and DynamicME within a unified platform, the application provides a practical and accessible entry point into proteome-constrained modeling.

While the underlying ME frameworks remain computationally challenging, the web-based design offloads this complexity to a remote worker system and returns organized, interpretable outputs. This structure makes it possible for both new and experienced users to explore metabolic knockouts, stress dependent proteome allocation, and dynamic growth simulations directly from their browser. Beyond accessibility, the reproducible architecture ensures that each simulation can be precisely replicated, supporting transparent and shareable research workflows.

By lowering the technical barriers to ME modeling, the COBRAme Web Application helps extend proteome-constrained approaches to a broader community of biologists, educators, and engineers. As additional ME frameworks are developed and standardized, the application can serve as a foundation for deploying future organism specific and multi-omics integrated models in a scalable, browser-based environment.

5.5 Acknowledgements

Funding for this research was provided by the Novo Nordisk Foundation through the Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517 and NNF20CC0035580). We would like to thank Marc Abrams for his assistance with manuscript editing. Special thanks to Dan Zielinski, Patrick Phaneuf, and Juan Tobicha-Bonilla for their feedback during development.

A.P., E.A.C., and B.O.P. designed and implemented the application. A.P. and B.O.P. wrote the manuscript, with contributions from all the other co-authors. All authors read and approved the final manuscript.

Chapter 5 in part, is currently being prepared for submission for publication.

Patel, Arjun; Catoiu, Edward A; Palsson, Bernhard O (2025). *COBRAme.org: A web application for solving and constraining metabolism and expression models.*

The dissertation author was the primary author.

Chapter 6. Conclusions

Cellular physiology is ultimately governed by the allocation of finite proteomic resources. Every metabolic flux, regulatory response, and phenotypic shift reflects tradeoffs in how the proteome is partitioned between competing cellular objectives. This principle has been recognized experimentally through global proteomics and transcriptomics, and it forms the conceptual foundation of this dissertation. Across each study, proteome allocation emerged not only as a measurable property but as the mechanistic driver of growth rate, stress response, and metabolic efficiency. The work presented here demonstrates that when mechanistic constraints on protein synthesis and enzyme usage are incorporated into computational models, the resulting simulations no longer reproduce cellular behavior merely by calibration, they recapitulate the underlying logic of resource allocation that defines microbial life.

Constraint-based modeling provides a natural language for expressing this logic. By imposing mass balance, thermodynamic feasibility, and enzyme capacity limits, these models formalize the boundary conditions under which evolution operates. Biological systems can then be understood as optimization problems constrained by physical and molecular laws, rather than as collections of independent reactions. This formulation has proven essential to connecting large-scale omics data to mechanistic prediction. Genome-scale metabolic reconstructions, when extended to include transcription, translation, and complex assembly, enable proteome allocation to emerge as an output of the optimization rather than as an empirical input. Within this framework, the proteome becomes the currency of cellular economy, linking gene expression, energy balance, and growth.

Each component of this dissertation illustrates a different aspect of this principle. Mechanistic simulations of bacterial metabolism and expression were used to interpret proteome tradeoffs under environmental and regulatory perturbations, where emergent sectoral shifts in the

proteome mirrored those observed experimentally. The integration of transcriptomic and proteomic constraints refined these predictions further, not by forcing agreement with data but by revealing which constraints most strongly shape feasible physiological states. In this way, omics integration is not merely a method of validation but a means of discovering emergent behavior. When high-dimensional data are projected onto mechanistic reconstructions, they expose the organizing rules of the cell, its resource limitations, co-regulatory structure, and hidden cost of unused capacity.

This emphasis on integrating experimental data within mechanistic frameworks aligns with a broader transformation in systems biology. The modern data landscape (transcriptomes, proteomes, metabolomes) provides unprecedented detail and can be produced at scale, but without a unifying model, such data cannot yield causal understanding. ME modeling resolves this gap by situating molecular measurements within the constraint structure of the cell. By coupling gene expression with metabolism through enzyme synthesis reactions, these models directly translate molecular level information into physiological predictions. The success of this approach depends on accurate biochemical reconstruction and on the ability to incorporate data in a way that respects biological mechanisms. The case studies presented here show that when this is achieved, emergent properties arise that were not explicitly encoded in the model: patterns of proteome reallocation, tradeoffs between biosynthetic and stress sectors, and shifts in respiratory efficiency that mirror experimentally observed growth laws.

Developing computational frameworks that make these models more accessible was a central goal of this work. The web application ecosystem built around COBRAme demonstrates how genome-scale ME models can be deployed, modified, and interrogated interactively. By integrating curated model repositories, editable parameters, and options for omics-based constraints, this system enables researchers to test hypotheses, visualize proteome allocations, and explore emergent predictions without requiring local installations or specialized code. The purpose

of this effort is not to propose a new theoretical model, but to lower the barrier between data and mechanistic interpretation. Also, to make proteome-constrained modeling a routine component of biological analysis. Through this accessibility, the mechanistic insights of ME modeling can reach the broader experimental community, accelerating iterative refinement between experiment and computation.

The comparative landscape of proteome-constrained frameworks further clarifies the advantages of ME models. GECKO, MOMENT, ETFL, and RBA each extend flux balance analysis by incorporating enzyme capacity or allocation constraints, yet they differ in how directly they connect molecular mechanisms to metabolic fluxes. GECKO imposes catalytic limits through measured or predicted turnover numbers and protein costs, effectively bounding reaction rates by enzyme availability (225). This formulation captures the concept of a proteome budget but treats translation as an external process. MOMENT, similarly, constrains total enzyme mass using kinetic parameters to predict growth rates and flux distributions, offering improved quantitative agreement but limited mechanistic depth (226). RBA abstracts the cell as an optimal resource allocation problem between functional protein sectors (227). Although individual enzymes and macromolecules are represented, their synthesis is modeled at the level of allocation and not through explicit gene expression reactions. Consequently, RBA lacks the direct gene-protein-reaction coupling that defines ME models. ETFL integrates expression and thermodynamic constraints into a mixed-integer linear framework, but still approximates the coupling between transcription, translation, and metabolism (228). In contrast, ME models implement this coupling explicitly: each enzymatic reaction is tied to its gene, transcript, and protein through defined synthesis and degradation processes. The resulting system retains linearity while enforcing a mechanistic relationship between expression and metabolism. This distinction is critical, the model

is not tuned to data, but rather derives phenotype from first principles constrained by biochemical reality.

Mechanistic coupling confers both predictive power and interpretability. Because each enzyme, ribosome, and metabolic complex is represented explicitly, changes in flux distribution can be traced directly to shifts in gene expression or protein synthesis costs. This level of detail allows ME models to connect multi-omics data to emergent physiological patterns in a way that abstracted models cannot. While frameworks such as GECKO or ETFL excel at large-scale parameterization and computational efficiency, they remain limited to describing how proteome allocation constrains metabolism. ME models, by contrast, can explain why those constraints exist, because the processes that generate them, transcription, translation, assembly, are part of the model itself. The distinction is one of mechanistic depth: ME models are not only enzyme-constrained but expression-driven, allowing constraint-based optimization to operate over the entire gene-to-flux axis of the cell.

Throughout this dissertation, proteome allocation has been shown to underlie the emergence of diverse bacterial phenotypes. When environmental conditions or regulatory programs change, the feasible space of proteome configurations shifts accordingly, and growth, stress resistance, and metabolic efficiency emerge as consequences of that redistribution. These are not imposed outcomes but emergent properties of the system's constraints.

Taken together, the studies in this dissertation demonstrate how the combination of mechanistic reconstruction, constraint-based optimization, and data integration can reveal the governing principles of cellular physiology. Proteome allocation provides the link between molecular biology and systems-level behavior, unifying growth, regulation, and adaptation within a single quantitative framework. The ME modeling formalism captures this relationship with the highest degree of mechanistic fidelity currently achievable, while the tools developed here make it

practical for broad application. Constraint-based modeling, at its core, mirrors the structure of biological reasoning: the cell is defined by constraints, driven by internal mechanisms, and shaped by optimization toward growth and survival. When those principles are encoded computationally and informed by experimental data, the resulting models do more than simulate metabolism, they expose the logic of life itself as an emergent consequence of constraint and expression.

References

1. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol. 1994 Oct;60(10):3724–31.

2. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A. 2000 May 9;97(10):5528–33.

3. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007 June 26;3:121.

4. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. 2003;4(9):R54.

5. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol Syst Biol. 2013 Oct 1;9:693.

6. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. PLoS Comput Biol. 2018 July;14(7):e1006302.

7. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Mishra AK, Das G, Malonia SK. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 2023 July 13;12(7):997.

8. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523–31.

9. Karp PD, Paley S, Caspi R, Kothari A, Krummenacker M, Midford PE, Moore LR, Subhraveti P, Gama-Castro S, Tierrafria VH, Lara P, Muñiz-Rascado L, Bonavides-Martinez C, Santos-Zavaleta A, Mackie A, Sun G, Ahn-Horst TA, Choi H, Covert MW, Collado-Vides J, Paulsen I. The EcoCyc Database (2023). EcoSal Plus. 2023 May 11;11(1):eesp-0002-2023.

10. Salgado H, Gama-Castro S, Lara P, Mejia-Almonte C, Alarcón-Carranza G, López-Almazo AG, Betancourt-Figueroa F, Peña-Loredo P, Alquicira-Hernández S, Ledezma-Tejeida D, Arizmendi-Zagal L, Mendez-Hernandez F, Diaz-Gomez AK, Ochoa-Praxedis E, Muñiz-Rascado LJ, García-Sotelo JS, Flores-Gallegos FA, Gómez L, Bonavides-Martínez C, del Moral-Chávez VM, Hernández-Alvarez AJ, Santos-Zavaleta A, Capella-Gutierrez S, Gelpi JL, Collado-Vides J. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. Nucleic Acids Res. 2023 Nov 16;52(D1):D255–64.

11. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

12. Catoiu EA, Krishnan J, Li G, Lou XA, Rychel K, Yuan Y, Bajpe H, Patel A, Choe D, Shin J, Burrows J, Phaneuf PV, Zielinski DC, Palsson BO. iModulonDB 2.0: dynamic tools to facilitate knowledge-mining and user-enabled analyses of curated transcriptomic datasets. Nucleic Acids Research. 2025 Jan 6;53(D1):D99–106.

13. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, Choudhary KS, Yang L, King ZA, Palsson BO. The Escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun. 2019 Dec 4;10(1):5536.

14. Lamoureux CR, Decker KT, Sastry AV, Rychel K, Gao Y, McConn JL, Zielinski DC, Palsson BO. A multi-scale expression and regulation knowledge base for Escherichia coli. Nucleic Acids Res. 2023 Oct 27;51(19):10176–93.

15. O'Brien EJ, Utrilla J, Palsson BO. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. PLoS Comput Biol. 2016 June;12(6):e1004998.

16. Yang L, Yurkovich JT, Lloyd CJ, Ebrahim A, Saunders MA, Palsson BO. Principles of proteome allocation are revealed using proteomic data and genome-scale models. Sci Rep. 2016 Nov 18;6:36734.

17. Yeung ES. Genome-wide correlation between mRNA and protein in a single cell. Angew Chem Int Ed Engl. 2011 Jan 17;50(3):583–5.

18. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 1999 Mar;19(3):1720–30.

19. Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr Genomics. 2013 Apr;14(2):91–110.

20. Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, Lerman JA, Lechner A, Sastry A, Bordbar A, Feist AM, Palsson BO. Multi-omic data integration enables discovery of hidden biological regularities. Nat Commun. 2016 Oct 26;7:13091.

21. Lahtvee PJ, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, Nielsen J. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst. 2017 May 24;4(5):495-504.e5.

22. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A, Sjöstedt E, Lundberg E, Szigyarto CAK, Skogs M, Takanen JO, Berling H, Tegel H, Mulder J, Nilsson P, Schwenk JM, Lindskog C, Danielsson F, Mardinoglu A, Sivertsson A, von Feilitzen K, Forsberg M, Zwahlen M, Olsson I, Navani S, Huss M, Nielsen J, Ponten F, Uhlén M. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014 Feb;13(2):397–406.

23. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018 Mar 15;9(1):1090.

24. Rychel K, Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. Nucleic Acids Res. 2021 Jan 8;49(D1):D112–20.

25. Sastry AV, Hu A, Heckmann D, Poudel S, Kavvas E, Palsson BO. Independent component analysis recovers consistent regulatory signals from disparate datasets. PLoS Comput Biol. 2021 Feb;17(2):e1008647.

26. Chauhan SM, Poudel S, Rychel K, Lamoureux C, Yoo R, Al Bulushi T, Yuan Y, Palsson BO, Sastry AV. Machine Learning Uncovers a Data-Driven Transcriptional Regulatory Network for the Crenarchaeal Thermoacidophile Sulfolobus acidocaldarius. Front Microbiol. 2021;12:753521.

27. Lim HG, Rychel K, Sastry AV, Bentley GJ, Mueller J, Schindel HS, Larsen PE, Laible PD, Guss AM, Niu W, Johnson CW, Beckham GT, Feist AM, Palsson BO. Machine-learning from Pseudomonas putida KT2440 transcriptomes reveals its transcriptional regulatory network. Metab Eng. 2022 July;72:297–310.

28. Rychel K, Sastry AV, Palsson BO. Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. Nat Commun. 2020 Dec 11;11(1):6338.

29. Yoo R, Rychel K, Poudel S, Al-Bulushi T, Yuan Y, Chauhan S, Lamoureux C, Palsson BO, Sastry A. Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. mSphere. 2022 Apr 27;7(2):e0003322.

30. Yuan Y, Seif Y, Rychel K, Yoo R, Chauhan S, Poudel S, Al-Bulushi T, Palsson BO, Sastry AV. Pan-Genome Analysis of Transcriptional Regulation in Six Salmonella enterica Serovar Typhimurium Strains Reveals Their Different Regulatory Structures. mSystems. 2022 Dec 20;7(6):e0046722.

31. Poudel S, Tsunemoto H, Seif Y, Sastry AV, Szubin R, Xu S, Machado H, Olson CA, Anand A, Pogliano J, Nizet V, Palsson BO. Revealing 29 sets of independently modulated genes in Staphylococcus aureus, their regulators, and role in key physiological response. Proc Natl Acad Sci U S A. 2020 July 21;117(29):17228–39.

32. Rajput A, Tsunemoto H, Sastry AV, Szubin R, Rychel K, Sugie J, Pogliano J, Palsson BO. Machine learning from Pseudomonas aeruginosa transcriptomes identifies independently modulated sets of genes associated with known transcriptional regulators. Nucleic Acids Res. 2022 Apr 22;50(7):3658–72.

33. Rodionova IA, Gao Y, Sastry A, Hefner Y, Lim HG, Rodionov DA, Saier MH, Palsson BO. Identification of a transcription factor, PunR, that regulates the purine and purine nucleoside transporter punC in E. coli. Commun Biol. 2021 Aug 19;4(1):991.

34. Anand A, Olson CA, Sastry AV, Patel A, Szubin R, Yang L, Feist AM, Palsson BO. Restoration of fitness lost due to dysregulation of the pyruvate dehydrogenase complex is triggered by ribosomal binding site modifications. Cell Rep. 2021 Apr 6;35(1):108961.

35. Anand A, Patel A, Chen K, Olson CA, Phaneuf PV, Lamoureux C, Hefner Y, Szubin R, Feist AM, Palsson BO. Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity. Nat Commun. 2022 June 27;13(1):3682.

36. Heckmann D, Campeau A, Lloyd CJ, Phaneuf PV, Hefner Y, Carrillo-Terrazas M, Feist AM, Gonzalez DJ, Palsson BO. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proc Natl Acad Sci U S A. 2020 Sept 15;117(37):23182–90.

37. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, Knoops K, Bauer M, Aebersold R, Heinemann M. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol. 2016 Jan;34(1):104–10.

38. Bathke J, Konzer A, Remes B, McIntosh M, Klug G. Comparative analyses of the variation of the transcriptome and proteome of Rhodobacter sphaeroides throughout growth. BMC Genomics. 2019 May 9;20(1):358.

39. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, Farber CR, Sinsheimer J, Kang HM, Furlotte N, Park CC, Wen PZ, Brewer H, Weitz K, Camp DG, Pan C, Yordanova R, Neuhaus I, Tilford C, Siemers N, Gargalovic P, Eskin E, Kirchgessner T, Smith DJ, Smith RD, Lusis AJ. Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 2011 June;7(6):e1001393.

40. Albrethsen J. Reproducibility in protein profiling by MALDI-TOF mass spectrometry. Clin Chem. 2007 May;53(5):852–8.

41. Comon P. Independent component analysis, A new concept? Signal Processing. 1994 Apr 1;36(3):287–314.

42. Weijland A, Harmark K, Cool RH, Anborgh PH, Parmeggiani A. Elongation factor Tu: a molecular switch in protein biosynthesis. Mol Microbiol. 1992 Mar;6(6):683–8.

43. González JC, Banerjee RV, Huang S, Sumner JS, Matthews RG. Comparison of cobalamin-independent and cobalamin-dependent methionine synthases from Escherichia coli: two solutions to the same chemical problem. Biochemistry. 1992 July 7;31(26):6045–56.

44. Mordukhova EA, Kim D, Pan JG. Stabilized homoserine o-succinyltransferases (MetA) or L-methionine partially recovers the growth defect in Escherichia coli lacking ATP-dependent proteases or the DnaK chaperone. BMC Microbiol. 2013 July 30;13:179.

45. Gold L. Posttranscriptional regulatory mechanisms in Escherichia coli. Annu Rev Biochem. 1988;57:199–233.

46. Yanofsky C. Attenuation in the control of expression of bacterial operons. Nature. 1981 Feb 26;289(5800):751–8.

47. Nudler E, Mironov AS. The riboswitch control of bacterial metabolism. Trends Biochem Sci. 2004 Jan;29(1):11–7.

48. Wessler SR, Calvo JM. Control of leu operon expression in Escherichia coli by a transcription attenuation mechanism. J Mol Biol. 1981 July 15;149(4):579–97.

49. Ontiveros-Palacios N, Smith AM, Grundy FJ, Soberon M, Henkin TM, Miranda-Ríos J. Molecular basis of gene regulation by the THI-box riboswitch. Mol Microbiol. 2008 Feb;67(4):793–803.

50. Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature. 2002 Oct 31;419(6910):952–6.

51. Giuliodori AM, Di Pietro F, Marzi S, Masquida B, Wagner R, Romby P, Gualerzi CO, Pon CL. The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. Mol Cell. 2010 Jan 15;37(1):21–33.

52. Yamanaka K, Inouye M. Selective mRNA degradation by polynucleotide phosphorylase in cold shock adaptation in Escherichia coli. J Bacteriol. 2001 May;183(9):2808–16.

53. Loewen PC, Hu B, Strutinsky J, Sparling R. Regulation in the rpoS regulon of Escherichia coli. Can J Microbiol. 1998 Aug;44(8):707–17.

54. Diez S, Ryu J, Caban K, Gonzalez RL, Dworkin J. The alarmones (p)ppGpp directly regulate translation initiation during entry into quiescence. Proc Natl Acad Sci U S A. 2020 July 7;117(27):15565–72.

55. Marincs F, Manfield IW, Stead JA, McDowall KJ, Stockley PG. Transcript analysis reveals an extended regulon and the importance of protein-protein co-operativity for the Escherichia coli methionine repressor. Biochem J. 2006 June 1;396(2):227–34.

56. Katz C, Rasouly A, Gur E, Shenhar Y, Biran D, Ron EZ. Temperature-dependent proteolysis as a control element in Escherichia coli metabolism. Res Microbiol. 2009 Nov;160(9):684–6.

57. Echandi G, Algranati ID. Defective 30S ribosomal particles in a polyamine auxotroph of Escherichia coli. Biochem Biophys Res Commun. 1975 Dec 1;67(3):1185–91.

58. Igarashi K, Kashiwagi K. Effects of polyamines on protein synthesis and growth of Escherichia coli. J Biol Chem. 2018 Nov 30;293(48):18702–9.

59. Hu XP, Schroeder S, Lercher MJ. Proteome efficiency of metabolic pathways in Escherichia coli increases along the nutrient flow. mSystems. 2023 Oct 26;8(5):e0076023.

60. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014 Apr 24;157(3):624–35.

61. Valgepea K, Peebo K, Adamberg K, Vilu R. Lean-proteome strains - next step in metabolic engineering. Front Bioeng Biotechnol. 2015;3:11.

62. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010 Nov 19;330(6007):1099–102.

63. Mori M, Zhang Z, Banaei-Esfahani A, Lalanne JB, Okano H, Collins BC, Schmidt A, Schubert OT, Lee DS, Li GW, Aebersold R, Hwa T, Ludwig C. From coarse to fine: the absolute Escherichia coli proteome under diverse growth conditions. Mol Syst Biol. 2021 May;17(5):e9536.

64. McConn JL, Lamoureux CR, Poudel S, Palsson BO, Sastry AV. Optimal dimensionality selection for independent component analysis of transcriptomic data. BMC Bioinformatics. 2021 Dec 8;22(1):584.

65. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc. 2007;2(8):1896–906.

66. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994 Nov;5(11):976–89.

67. Ahrné E, Molzahn L, Glatter T, Schmidt A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics. 2013 Sept;13(17):2567–78.

68. Neidhardt FC, editor. Escherichia coli and Salmonella: cellular and molecular biology. 2nd ed. Washington, D.C: ASM Press; 1996. 2 p.

69. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Evolution of Electron-Transport Chains. In: Molecular Biology of the Cell 4th edition [Internet]. Garland Science; 2002 [cited 2025 July 31]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26849/

70. Brochier-Armanet C, Talla E, Gribaldo S. The multiple evolutionary histories of dioxygen reductases: Implications for the origin and evolution of aerobic respiration. Mol Biol Evol. 2009 Feb;26(2):285–97.

71. Sturm G, Richter K, Doetsch A, Heide H, Louro RO, Gescher J. A dynamic periplasmic electron transfer network enables respiratory flexibility beyond a thermodynamic regulatory regime. ISME J. 2015 Aug;9(8):1802–11.

72. Chen K, Anand A, Olson C, Sandberg TE, Gao Y, Mih N, Palsson BO. Bacterial fitness landscapes stratify based on proteome allocation associated with discrete aero-types. PLoS Comput Biol. 2021 Jan;17(1):e1008596.

73. Unden G, Bongaerts J. Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors. Biochim Biophys Acta. 1997 July 4;1320(3):217–34.

74. Borisov VB, Murali R, Verkhovskaya ML, Bloch DA, Han H, Gennis RB, Verkhovsky MI. Aerobic respiratory chain of Escherichia coli is not allowed to work in fully uncoupled mode. Proc Natl Acad Sci U S A. 2011 Oct 18;108(42):17320–4.

75. Ingledew WJ, Poole RK. The respiratory chains of Escherichia coli. Microbiol Rev. 1984 Sept;48(3):222–71.

76. Phaneuf PV, Gosting D, Palsson BO, Feist AM. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. Nucleic Acids Res. 2019 Jan 8;47(D1):D1164–71.

77. LaCroix RA, Sandberg TE, O'Brien EJ, Utrilla J, Ebrahim A, Guzman GI, Szubin R, Palsson BO, Feist AM. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of Escherichia coli K-12 MG1655 on glucose minimal medium. Appl Environ Microbiol. 2015 Jan;81(1):17–30.

78. Sandberg TE, Lloyd CJ, Palsson BO, Feist AM. Laboratory Evolution to Alternating Substrate Environments Yields Distinct Phenotypic and Genetic Adaptive Strategies. Appl Environ Microbiol. 2017 July 1;83(13):e00410-17.

79. Lennen RM, Jensen K, Mohammed ET, Malla S, Börner RA, Chekina K, Özdemir E, Bonde I, Koza A, Maury J, Pedersen LE, Schöning LY, Sonnenschein N, Palsson BO, Sommer MOA, Feist AM, Nielsen AT, Herrgård MJ. Adaptive laboratory evolution reveals general and specific chemical tolerance mechanisms and enhances biochemical production [Internet]. bioRxiv; 2019 [cited 2025 July 31]. p. 634105. Available from: https://www.biorxiv.org/content/10.1101/634105v1

80. Sandberg TE, Salazar MJ, Weng LL, Palsson BO, Feist AM. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. Metab Eng. 2019 Dec;56:1–16.

81. Acharya S, Foster PL, Brooks P, Fishel R. The coordinated functions of the E. coli MutS and MutL proteins in mismatch repair. Mol Cell. 2003 July;12(1):233–46.

82. González-González A, Hug SM, Rodríguez-Verdugo A, Patel JS, Gaut BS. Adaptive Mutations in RNA Polymerase and the Transcriptional Terminator Rho Have Similar Effects on Escherichia coli Gene Expression. Mol Biol Evol. 2017 Nov 1;34(11):2839–55.

83. Conrad TM, Frazier M, Joyce AR, Cho BK, Knight EM, Lewis NE, Landick R, Palsson BØ. RNA polymerase mutants found through adaptive evolution reprogram Escherichia coli for optimal growth in minimal media. Proc Natl Acad Sci U S A. 2010 Nov 23;107(47):20500–5.

84. Long A, Liti G, Luptak A, Tenaillon O. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nat Rev Genet. 2015 Oct;16(10):567–82.

85. Utrilla J, O'Brien EJ, Chen K, McCloskey D, Cheung J, Wang H, Armenta-Medina D, Feist AM, Palsson BO. Global Rebalancing of Cellular Resources by Pleiotropic Point Mutations Illustrates a Multi-scale Mechanism of Adaptive Evolution. Cell Syst. 2016 Apr 27;2(4):260–71.

86. Thomas AK, Preetha S, Omanakuttan A, Vidyullata L, Ashokan A, Rajachandran V, Chattopadhyay S. Mutational convergence acts as a major player in adaptive parallel evolution of Shigella spp. Sci Rep. 2019 Mar 1;9(1):3252.

87. Horinouchi T, Suzuki S, Hirasawa T, Ono N, Yomo T, Shimizu H, Furusawa C. Phenotypic convergence in bacterial adaptive evolution to ethanol stress. BMC Evol Biol. 2015 Sept 3;15:180.

88. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016 Jan;11(1):1–9.

89. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013 Jan;10(1):47–53.

90. Steinsiek S, Frixel S, Stagge S, SUMO, Bettenbrock K. Characterization of E. coli MG1655 and frdA and sdhC mutants at various aerobiosis levels. J Biotechnol. 2011 June 10;154(1):35–45.

91. Zheng J, Singh VK, Jia Z. Identification of an ITPase/XTPase in Escherichia coli by structural and biochemical analysis. Structure. 2005 Oct;13(10):1511–20.

92. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering C von. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019 Jan 8;47(D1):D607–13.

93. Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res. 2017 Apr 20;45(7):3615–26.

94. Chen K, Gao Y, Mih N, O'Brien EJ, Yang L, Palsson BO. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc Natl Acad Sci U S A. 2017 Oct 24;114(43):11548–53.

95. Price CE, Driessen AJM. Biogenesis of membrane bound respiratory complexes in Escherichia coli. Biochim Biophys Acta. 2010 June;1803(6):748–66.

96. Young IG, Jaworowski A, Poulis MI. Amplification of the respiratory NADH dehydrogenase of Escherichia coli by gene cloning. Gene. 1978 Sept;4(1):25–36.

97. Vamshi Krishna K, Venkata Mohan S. Purification and Characterization of NDH-2 Protein and Elucidating Its Role in Extracellular Electron Transport and Bioelectrogenic Activity. Front Microbiol. 2019;10:880.

98. Stettner AI, Segrè D. The cost of efficiency in energy metabolism. Proc Natl Acad Sci U S A. 2013 June 11;110(24):9629–30.

99. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R. Glycolytic strategy as a tradeoff between energy yield and protein cost. Proc Natl Acad Sci U S A. 2013 June 11;110(24):10039–44.

100. Ferguson SJ. ATP synthase: from sequence to ring size to the P/O ratio. Proc Natl Acad Sci U S A. 2010 Sept 28;107(39):16755–6.

101. Kaila VRI, Wikström M. Architecture of bacterial respiratory chains. Nat Rev Microbiol. 2021 May;19(5):319–30.

102. Sobti M, Walshe JL, Wu D, Ishmukhametov R, Zeng YC, Robinson CV, Berry RM, Stewart AG. Cryo-EM structures provide insight into how E. coli F1Fo ATP synthase accommodates symmetry mismatch. Nat Commun. 2020 May 26;11(1):2615.

103. Petersen J, Förster K, Turina P, Gräber P. Comparison of the H+/ATP ratios of the H+-ATP synthases from yeast and from chloroplast. Proc Natl Acad Sci U S A. 2012 July 10;109(28):11150–5.

104. Steigmiller S, Turina P, Gräber P. The thermodynamic H+/ATP ratios of the H+-ATPsynthases from chloroplasts and Escherichia coli. Proc Natl Acad Sci U S A. 2008 Mar 11;105(10):3745–50.

105. Jiang W, Hermolin J, Fillingame RH. The preferred stoichiometry of c subunits in the rotary motor sector of Escherichia coli ATP synthase is 10. Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):4966–71.

106. Preiss L, Klyszejko AL, Hicks DB, Liu J, Fackelmayer OJ, Yildiz Ö, Krulwich TA, Meier T. The c-ring stoichiometry of ATP synthase is adapted to cell physiological requirements of alkaliphilic Bacillus pseudofirmus OF4. Proc Natl Acad Sci U S A. 2013 May 7;110(19):7874–9.

107. Schemidt RA, Qu J, Williams JR, Brusilow WS. Effects of carbon source on expression of F0 genes and on the stoichiometry of the c subunit in the F1F0 ATPase of Escherichia coli. J Bacteriol. 1998 June;180(12):3205–8.

108. Tomashek JJ, Brusilow WS. Stoichiometry of energy coupling by proton-translocating ATPases: a history of variability. J Bioenerg Biomembr. 2000 Oct;32(5):493–500.

109. Thomason LC, Costantino N, Court DL. E. coli Genome Manipulation by P1 Transduction. Current Protocols in Molecular Biology. 2007;79(1):1.17.1-1.17.8.

110. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology. 2006 Jan;2(1):2006.0008.

111. Ou J, Zhu LJ. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. Nat Methods. 2019 June;16(6):453–4.

112. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018 Nov 1;34(21):3755–8.

113. Wagih O, Galardini M, Busby BP, Memon D, Typas A, Beltrao P. A resource of variant effect predictions of single nucleotide variants in model organisms. Mol Syst Biol. 2018 Dec 20;14(12):e8430.

114. Anand A, Olson CA, Yang L, Sastry AV, Catoiu E, Choudhary KS, Phaneuf PV, Sandberg TE, Xu S, Hefner Y, Szubin R, Feist AM, Palsson BO. Pseudogene repair driven by selection pressure applied in experimental evolution. Nat Microbiol. 2019 Mar;4(3):386–9.

115. Srivastav S, Biswas A, Anand A. Interplay of niche and respiratory network in shaping bacterial colonization. Journal of Biological Chemistry. 2025 Jan 1;301(1):108052.

116. Farr SB, Kogoma T. Oxidative stress responses in Escherichia coli and Salmonella typhimurium. Microbiological Reviews. 1991 Dec;55(4):561–85.

117. Nyström T. The glucose-starvation stimulon of Escherichia coli: induced and repressed synthesis of enzymes of central metabolic pathways and role of acetyl phosphate in gene expression and starvation survival. Mol Microbiol. 1994 June;12(5):833–43.

118. Minagawa S, Ogasawara H, Kato A, Yamamoto K, Eguchi Y, Oshima T, Mori H, Ishihama A, Utsumi R. Identification and Molecular Characterization of the Mg2+ Stimulon of Escherichia coli. Journal of Bacteriology. 2003 July;185(13):3696–702.

119. Schumann W. Regulation of bacterial heat shock stimulons. Cell Stress Chaperones. 2016 Nov;21(6):959–68.

120. Grifantini R, Toukoki C, Colaprico A, Gryllos I. Peroxide Stimulon and Role of PerR in Group A Streptococcus. Journal of Bacteriology. 2011 Nov 9;193(23):6539–51.

121. Song Y, Lunde CS, Benton BM, Wilkinson BJ. Further Insights into the Mode of Action of the Lipoglycopeptide Telavancin through Global Gene Expression Studies. Antimicrobial Agents and Chemotherapy. 2012 May 13;56(6):3157–64.

122. Xiao Q, Vakulenko S, Chang M, Mobashery S. Mutations in mmpL and in the Cell Wall Stress Stimulon Contribute to Resistance to Oxadiazole Antibiotics in Methicillin-Resistant Staphylococcus aureus. Antimicrobial Agents and Chemotherapy. 2014 Sept 15;58(10):5841–7.

123. Hara KY, Shimodate N, Ito M, Baba T, Mori H, Mori H. Systematic genome-wide scanning for genes involved in ATP generation in *Escherichia coli*. Metabolic Engineering. 2009 Jan 1;11(1):1–7.

124. Wang X, Xia K, Yang X, Tang C. Growth strategy of microbes on mixed carbon sources. Nat Commun. 2019 Mar 20;10(1):1279.

125. Skotnicki ML, Rolfe BG. Pathways of energy metabolism required for phenotypic expression of nif+Kp genes in Escherichia coli. Aust J Biol Sci. 1979 Dec;32(6):637–49.

126. Martínez-Gómez K, Flores N, Castañeda HM, Martínez-Batallar G, Hernández-Chávez G, Ramírez OT, Gosset G, Encarnación S, Bolivar F. New insights into Escherichia coli metabolism: carbon scavenging, acetate metabolism and carbon recycling responses during growth on glycerol. Microb Cell Fact. 2012 Apr 18;11(1):46.

127. O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular Systems Biology. 2013 Jan;9(1):693.

128. Thiele I, Jamshidi N, Fleming RMT, Palsson BØ. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its

mathematical formulation, and its functional characterization. PLoS Comput Biol. 2009 Mar;5(3):e1000312.

129. Nitzschke A, Bettenbrock K. All three quinone species play distinct roles in ensuring optimal growth under aerobic and fermentative conditions in E. coli K12. PLoS One. 2018;13(4):e0194699.

130. Benoit S, Abaibou H, Mandrand-Berthelot MA. Topological Analysis of the Aerobic Membrane-Bound Formate Dehydrogenase of Escherichia coli. Journal of Bacteriology. 1998 Dec 15;180(24):6625–34.

131. Abaibou H, Pommier J, Benoit S, Giordano G, Mandrand-Berthelot MA. Expression and characterization of the Escherichia coli fdo locus and a possible physiological role for aerobic formate dehydrogenase. Journal of Bacteriology. 1995 Dec;177(24):7141–9.

132. Patel A, McGrosso D, Hefner Y, Campeau A, Sastry AV, Maurya S, Rychel K, Gonzalez DJ, Palsson BO. Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. Nat Commun. 2024 June 19;15(1):5234.

133. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ. The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS Genet. 2013;9(10):e1003839.

134. Cunin R, Glansdorff N, Piérard A, Stalon V. Biosynthesis and metabolism of arginine in bacteria. Microbiol Rev. 1986 Sept;50(3):314–52.

135. Cronan JE. The Escherichia coli FadR transcription factor: Too much of a good thing? Mol Microbiol. 2021 June;115(6):1080–5.

136. Metzner M, Germer J, Hengge R. Multiple stress signal integration in the regulation of the complex sigma S-dependent csiD-ygaF-gabDTP operon in Escherichia coli. Mol Microbiol. 2004 Feb;51(3):799–811.

137. Pellicer MT, Badía J, Aguilar J, Baldomà L. glc locus of Escherichia coli: characterization of genes encoding the subunits of glycolate oxidase and the glc regulator protein. Journal of Bacteriology. 1996 Apr;178(7):2051–9.

138. Nemoto N, Kurihara S, Kitahara Y, Asada K, Kato K, Suzuki H. Mechanism for Regulation of the Putrescine Utilization Pathway by the Transcription Factor PuuR in Escherichia coli K-12. Journal of Bacteriology. 2012 June 11;194(13):3437–47.

139. Choudhury A, Gachet B, Dixit Z, Faure R, Gill RT, Tenaillon O. Deep mutational scanning reveals the molecular determinants of RNA polymerase-mediated adaptation and tradeoffs. Nat Commun. 2023 Oct 9;14(1):6319.

140. Janausch IG, Zientz E, Tran QH, Kröger A, Unden G. C4-dicarboxylate carriers and sensors in bacteria. Biochim Biophys Acta. 2002 Jan 17;1553(1–2):39–56.

141. Boogerd FC, Boe L, Michelsen O, Jensen PR. atp Mutants of Escherichia coli Fail To Grow on Succinate Due to a Transport Deficiency. Journal of Bacteriology. 1998 Nov 15;180(22):5855–9.

142. Applebee MK, Joyce AR, Conrad TM, Pettigrew DW, Palsson BØ. Functional and Metabolic Effects of Adaptive Glycerol Kinase (GLPK) Mutants in Escherichia coli*. Journal of Biological Chemistry. 2011 July 1;286(26):23150–9.

143. Anand A, Chen K, Yang L, Sastry AV, Olson CA, Poudel S, Seif Y, Hefner Y, Phaneuf PV, Xu S, Szubin R, Feist AM, Palsson BO. Adaptive evolution reveals a tradeoff between growth rate and oxidative stress during naphthoquinone-based aerobic respiration. Proceedings of the National Academy of Sciences. 2019 Dec 10;116(50):25287–92.

144. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016 Jan 4;44(D1):D515-522.

145. Yang L, Ma D, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO. solveME: fast and reliable solution of nonlinear ME models. BMC Bioinformatics. 2016 Sept 22;17(1):391.

146. Sastry AV, Yuan Y, Poudel S, Rychel K, Yoo R, Lamoureux CR, Li G, Burrows JT, Chauhan S, Haiman ZB, Al Bulushi T, Seif Y, Palsson BO, Zielinski DC. iModulonMiner and PyModulon: Software for unsupervised mining of gene expression compendia. PLoS Comput Biol. 2024 Oct;20(10):e1012546.

147. Goelzer A, Fromion V. Bacterial growth rate reflects a bottleneck in resource allocation. Biochimica et Biophysica Acta (BBA) - General Subjects. 2011 Oct 1;1810(10):978–88.

148. Irving SE, Choudhury NR, Corrigan RM. The stringent response and physiological roles of (pp)pGpp in bacteria. Nat Rev Microbiol. 2021 Apr;19(4):256–71.

149. Gottesman S. Trouble is coming: Signaling pathways that regulate general stress responses in bacteria. J Biol Chem. 2019 Aug 2;294(31):11685–700.

150. Schellhorn HE. Function, Evolution, and Composition of the RpoS Regulon in Escherichia coli. Front Microbiol. 2020;11:560099.

151. Anand A, Chen K, Catoiu E, Sastry AV, Olson CA, Sandberg TE, Seif Y, Xu S, Szubin R, Yang L, Feist AM, Palsson BO. OxyR Is a Convergent Target for Mutations Acquired during Adaptation to Oxidative Stress-Prone Metabolic States. Mol Biol Evol. 2020 Mar 1;37(3):660–7.

152. Hidalgo D, Martínez-Ortiz CA, Palsson BO, Jiménez JI, Utrilla J. Regulatory perturbations of ribosome allocation in bacteria reshape the growth proteome with a trade-off in adaptation capacity. iScience. 2022 Mar 18;25(3):103879.

153. Tan J, Sastry AV, Fremming KS, Bjørn SP, Hoffmeyer A, Seo S, Voldborg BG, Palsson BO. Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. Metab Eng. 2020 Sept;61:360–8.

154. Sastry A, Dillon N, Poudel S, Hefner Y, Xu S, Szubin R, Feist A, Nizet V, Palsson B. Decomposition of transcriptional responses provides insights into differential antibiotic susceptibility [Internet]. bioRxiv; 2020 [cited 2025 Oct 14]. p. 2020.05.04.077271. Available from: https://www.biorxiv.org/content/10.1101/2020.05.04.077271v1

155. Wang D, Bushnell DA, Westover KD, Kaplan CD, Kornberg RD. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell. 2006 Dec 1;127(5):941–54.

156. Jovanovic M, Burrows PC, Bose D, Cámara B, Wiesler S, Zhang X, Wigneshweraraj S, Weinzierl ROJ, Buck M. Activity map of the Escherichia coli RNA polymerase bridge helix. J Biol Chem. 2011 Apr 22;286(16):14469–79.

157. Ederth J, Artsimovitch I, Isaksson LA, Landick R. The downstream DNA jaw of bacterial RNA polymerase facilitates both transcriptional initiation and pausing. J Biol Chem. 2002 Oct 4;277(40):37456–63.

158. Ross W, Sanchez-Vazquez P, Chen AY, Lee JH, Burgos HL, Gourse RL. ppGpp Binding to a Site at the RNAP-DksA Interface Accounts for Its Dramatic Effects on Transcription Initiation during the Stringent Response. Mol Cell. 2016 June 16;62(6):811–23.

159. Deighan P, Diez CM, Leibman M, Hochschild A, Nickels BE. The bacteriophage lambda Q antiterminator protein contacts the beta-flap domain of RNA polymerase. Proc Natl Acad Sci U S A. 2008 Oct 7;105(40):15305–10.

160. Jin DJ, Cashel M, Friedman DI, Nakamura Y, Walter WA, Gross CA. Effects of rifampicin resistant rpoB mutations on antitermination and interaction with nusA in Escherichia coli. J Mol Biol. 1988 Nov 20;204(2):247–61.

161. Parshin A, Shiver AL, Lee J, Ozerova M, Schneidman-Duhovny D, Gross CA, Borukhov S. DksA regulates RNA polymerase in Escherichia coli through a network of interactions in the secondary channel that includes Sequence Insertion 1. Proc Natl Acad Sci U S A. 2015 Dec 15;112(50):E6862-6871.

162. Zhou Y, Zhang X, Ebright RH. Identification of the activating region of catabolite gene activator protein (CAP): isolation and characterization of mutants of CAP specifically defective in transcription activation. Proc Natl Acad Sci U S A. 1993 July 1;90(13):6081–5.

163. Rhodius VA, Busby SJ. Transcription activation by the Escherichia coli cyclic AMP receptor protein: determinants within activating region 3. J Mol Biol. 2000 June 2;299(2):295–310.

164. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics. 2010 Mar 1;26(5):689–91.

165. Dalldorf C, Rychel K, Szubin R, Hefner Y, Patel A, Zielinski DC, Palsson BO. The hallmarks of a tradeoff in transcriptomes that balances stress and growth functions. mSystems. 2024 June 3;9(7):e00305-24.

166. Chen K, Gao Y, Mih N, O'Brien EJ, Yang L, Palsson BO. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc Natl Acad Sci U S A. 2017 Oct 24;114(43):11548–53.

167. Hassan HM, Fridovich I. Intracellular production of superoxide radical and of hydrogen peroxide by redox active compounds. Arch Biochem Biophys. 1979 Sept;196(2):385–95.

168. Hassan HM, Fridovich I. Paraquat and Escherichia coli. Mechanism of production of extracellular superoxide radical. J Biol Chem. 1979 Nov 10;254(21):10846–52.

169. Imlay JA. Where in the world do bacteria experience oxidative stress? Environ Microbiol. 2019 Feb;21(2):521–30.

170. Peabody GL, Winkler J, Kao KC. Tools for developing tolerance to toxic chemicals in microbial systems and perspectives on moving the field forward and into the industrial setting. Current Opinion in Chemical Engineering. 2014 Nov 1;6:9–17.

171. Yang L, Mih N, Anand A, Park JH, Tan J, Yurkovich JT, Monk JM, Lloyd CJ, Sandberg TE, Seo SW, Kim D, Sastry AV, Phaneuf P, Gao Y, Broddrick JT, Chen K, Heckmann D, Szubin R, Hefner Y, Feist AM, Palsson BO. Cellular responses to reactive oxygen species are predicted from molecular mechanisms. Proc Natl Acad Sci U S A. 2019 July 9;116(28):14368–73.

172. Rychel K, Tan J, Patel A, Lamoureux C, Hefner Y, Szubin R, Johnsen J, Mohamed ETT, Phaneuf PV, Anand A, Olson CA, Park JH, Sastry AV, Yang L, Feist AM, Palsson BO. Laboratory evolution, transcriptomics, and modeling reveal mechanisms of paraquat tolerance. Cell Reports [Internet]. 2023 Sept 26 [cited 2025 Apr 22];42(9). Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(23)01116-6

173. Clomburg JM, Crumbley AM, Gonzalez R. Industrial biomanufacturing: The future of chemical production. Science. 2017 Jan 6;355(6320):aag0804.

174. Choi KR, Jang WD, Yang D, Cho JS, Park D, Lee SY. Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. Trends Biotechnol. 2019 Aug;37(8):817–37.

175. Shiloach J, Fass R. Growing E. coli to high cell density--a historical perspective on method development. Biotechnol Adv. 2005 July;23(5):345–57.

176. Lee SY. High cell-density culture of Escherichia coli. Trends Biotechnol. 1996 Mar;14(3):98–105.

177. Liu X, Ng C, Ferenci T. Global adaptations resulting from high population densities in Escherichia coli cultures. J Bacteriol. 2000 Aug;182(15):4158–64.

178. Beulig F, Bafna-Rührer J, Jensen PE, Kim SH, Patel A, Kandasamy V, Steffen C, Decker K, Zielinski D, Yang L, Özdemir E, Sudarsan S, Palsson B. Trade-off between resistance and persistence in high cell density cultures. mSystems. 2025 June 13;10(7):e00323-25.

179. Friedrich T, Pohl T. NADH as Donor. EcoSal Plus. 2007 Aug 13;2(2):10.1128/ecosalplus.3.2.4.

180. Unden G, Steinmetz PA, Degreif-Dünnwald P. The Aerobic and Anaerobic Respiratory Chain of Escherichia coli and Salmonella enterica: Enzymes and Energetics. EcoSal Plus. 2014 July 18;6(1):10.1128/ecosalplus.ESP-0005–2013.

181. Zerfaß C, Asally M, Soyer OS. Interrogating metabolism as an electron flow system. Current Opinion in Systems Biology. 2019 Feb 1;13:59–67.

182. Cole JA, Richardson DJ. Respiration of Nitrate and Nitrite. EcoSal Plus. 2008 Aug 18;3(1):10.1128/ecosal.3.2.5.

183. Gralnick JA, Bond DR. Electron Transfer Beyond the Outer Membrane: Putting Electrons to Rest. Annual Review of Microbiology. 2023 Sept 15;77(Volume 77, 2023):517–39.

184. Jeuken LJC, Hards K, Nakatani Y. Extracellular Electron Transfer: Respiratory or Nutrient Homeostasis? Journal of Bacteriology. 2020 Mar 11;202(7):10.1128/jb.00029-20.

185. Lovley DR, Phillips EJP. Novel Mode of Microbial Energy Metabolism: Organic Carbon Oxidation Coupled to Dissimilatory Reduction of Iron or Manganese. Applied and Environmental Microbiology. 1988 June;54(6):1472–80.

186. Myers CR, Nealson KH. Bacterial Manganese Reduction and Growth with Manganese Oxide as the Sole Electron Acceptor. Science. 1988 June 3;240(4857):1319–21.

187. Kundu BB, Krishnan J, Szubin R, Patel A, Palsson BO, Zielinski DC, Ajo-Franklin CM. Extracellular respiration is a latent energy metabolism in Escherichia coli. Cell. 2025 May 29;188(11):2907-2924.e23.

188. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, Feist AM, Palsson BO. iML1515, a knowledgebase that computes Escherichia coli traits. Nat Biotechnol. 2017 Oct;35(10):904–8.

189. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010 Mar;28(3):245–8.

190. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, Magnusdóttir S, Ng CY, Preciat G, Žagare A, Chan SHJ, Aurich MK, Clancy CM, Modamio J, Sauls JT, Noronha A, Bordbar A, Cousins B, El Assal DC, Valcarcel LV, Apaolaza I, Ghaderi S, Ahookhosh M, Ben Guebila M, Kostromins A, Sompairac N, Le HM, Ma D, Sun Y, Wang L, Yurkovich JT, Oliveira MAP, Vuong PT, El Assal LP, Kuperstein I, Zinovyev A, Hinton HS, Bryant WA, Aragón Artacho FJ, Planes FJ, Stalidzans E, Maass A, Vempala S, Hucka M, Saunders MA, Maranas CD, Lewis NE, Sauter T, Palsson BØ, Thiele I, Fleming RMT. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019 Mar;14(3):639–702.

191. Kolb A, Busby S, Buc H, Garges S, Adhya S. Transcriptional regulation by cAMP and its receptor protein. Annu Rev Biochem. 1993;62:749–95.

192. Potrykus K, Cashel M. (p)ppGpp: still magical? Annu Rev Microbiol. 2008;62:35–51.

193. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. Multidimensional optimality of microbial metabolism. Science. 2012 May 4;336(6081):601–4.

194. You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, Wang YP, Lenz P, Yan D, Hwa T. Coordination of bacterial proteome with metabolism by cyclic AMP signalling. Nature. 2013 Aug 15;500(7462):301–6.

195. Battesti A, Majdalani N, Gottesman S. The RpoS-mediated general stress response in Escherichia coli. Annu Rev Microbiol. 2011;65:189–213.

196. Kim D, Seo SW, Gao Y, Nam H, Guzman GI, Cho BK, Palsson BO. Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. Nucleic Acids Res. 2018 Apr 6;46(6):2901–17.

197. Latif H, Federowicz S, Ebrahim A, Tarasova J, Szubin R, Utrilla J, Zengler K, Palsson BO. ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions. PLOS ONE. 2018 May 17;13(5):e0197272.

198. Ghatak S, King ZA, Sastry A, Palsson BO. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. Nucleic Acids Res. 2019 Mar 18;47(5):2446–54.

199. Park JY, Lee SM, Ebrahim A, Scott-Nevros ZK, Kim J, Yang L, Sastry A, Seo SW, Palsson BO, Kim D. Model-driven experimental design workflow expands understanding of regulatory role of Nac in Escherichia coli. NAR Genom Bioinform. 2023 Mar;5(1):lqad006.

200. Lu CD, Abdelal AT. Role of ArgR in Activation of the ast Operon, Encoding Enzymes of the Arginine Succinyltransferase Pathway in Salmonella typhimurium. J Bacteriol. 1999 Mar;181(6):1934–8.

201. Dolan SK, Wijaya A, Geddis SM, Spring DR, Silva-Rocha R, Welch M. Loving the poison: the methylcitrate cycle and bacterial pathogenesis. Microbiology. 2018 Mar 1;164(3):251–9.

202. Gregersen N. The specific inhibition of the pyruvate dehydrogenase complex from pig kidney by propionyl-CoA and isovaleryl-CoA. Biochemical Medicine. 1981 Aug 1;26(1):20–7.

203. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016 June 20;17(1):132.

204. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021 Apr;18(4):366–8.

205. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep. 2013 Feb;5(1):58–65.

206. Greenberg DB, Stülke J, Saier MH. Domain analysis of transcriptional regulators bearing PTS regulatory domains. Res Microbiol. 2002 Oct;153(8):519–26.

207. Lempp M, Farke N, Kuntz M, Freibert SA, Lill R, Link H. Systematic identification of metabolites controlling gene expression in E. coli. Nat Commun. 2019 Oct 2;10(1):4463.

208. Anand A, Chen K, Catoiu E, Sastry AV, Olson CA, Sandberg TE, Seif Y, Xu S, Szubin R, Yang L, Feist AM, Palsson BO. OxyR Is a Convergent Target for Mutations Acquired during Adaptation to Oxidative Stress-Prone Metabolic States. Mol Biol Evol. 2020 Mar 1;37(3):660–7.

209. Bershtein S, Choi JM, Bhattacharyya S, Budnik B, Shakhnovich E. Systems-level response to point mutations in a core metabolic enzyme modulates genotype-phenotype relationship. Cell Rep. 2015 Apr 28;11(4):645–56.

210. Chen H, Bjerknes M, Kumar R, Jay E. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. Nucleic Acids Res. 1994 Nov 25;22(23):4953–7.

211. Mendoza-Chamizo B, Løbner-Olesen A, Charbon G. Coping with Reactive Oxygen Species to Ensure Genome Stability in Escherichia coli. Genes (Basel). 2018 Nov 21;9(11):565.

212. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010 Nov 19;330(6007):1099–102.

213. Scott M, Klumpp S, Mateescu EM, Hwa T. Emergence of robust growth laws from optimal regulation of ribosome synthesis. Mol Syst Biol. 2014 Aug 22;10(8):747.

214. O'Brien EJ, Monk JM, Palsson BO. Using Genome-Scale Models to Predict Biological Capabilities. Cell. 2015 May 21;161(5):971–87.

215. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology. 2019 Aug 7;20(1):158.

216. Passi A, Tibocha-Bonilla JD, Kumar M, Tec-Campos D, Zengler K, Zuniga C. Genome-Scale Metabolic Modeling Enables In-Depth Understanding of Big Data. Metabolites. 2021 Dec 24;12(1):14.

217. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet. 2014 Feb;15(2):107–20.

218. Dahal S, Zhao J, Yang L. Recent advances in genome-scale modeling of proteome allocation. Current Opinion in Systems Biology. 2021 June 1;26:39–45.

219. Tibocha-Bonilla JD, Zuñiga C, Lekbua A, Lloyd C, Rychel K, Short K, Zengler K. Predicting stress response and improved protein overproduction in Bacillus subtilis. NPJ Syst Biol Appl. 2022 Dec 27;8(1):50.

220. Liu JK, Lloyd C, Al-Bassam MM, Ebrahim A, Kim JN, Olson C, Aksenov A, Dorrestein P, Zengler K. Predicting proteome allocation, overflow metabolism, and metal requirements in a model acetogen. PLoS Comput Biol. 2019 Mar 7;15(3):e1006848.

221. Zhao J, Chen K, Palsson BO, Yang L. StressME: Unified computing framework of Escherichia coli metabolism, gene expression, and stress responses. PLOS Computational Biology. 2024 Feb 12;20(2):e1011865.

222. Yang L, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO. DynamicME: dynamic simulation and refinement of integrated models of metabolism and protein expression. BMC Systems Biology. 2019 Jan 9;13(1):2.

223. Du B, Yang L, Lloyd CJ, Fang X, Palsson BO. Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in Escherichia coli. PLOS Computational Biology. 2019 Dec 6;15(12):e1007525.

224. Patel A, Banwani N, Mink R, Prabhakaran DM, Khairnar SV, Feist AM, Palsson BO, Anand A. Aerobicity stimulon in Escherichia coli revealed using multi-scale computational systems biology of adapted respiratory variants [Internet]. bioRxiv; 2025 [cited 2025 Aug 25]. p. 2025.03.13.642450. Available from: https://www.biorxiv.org/content/10.1101/2025.03.13.642450v1

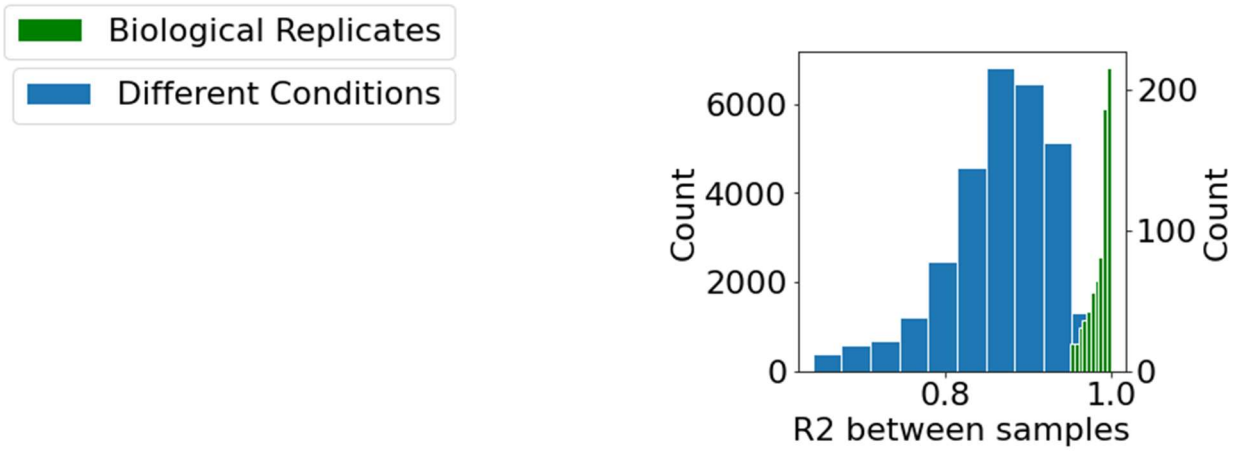
225. Chen Y, Gustafsson J, Tafur Rangel A, Anton M, Domenzain I, Kittikunapong C, Li F, Yuan L, Nielsen J, Kerkhoven EJ. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. Nat Protoc. 2024 Mar;19(3):629–67.

226. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol. 2012;8(7):e1002575.

227. Bulović A, Fischer S, Dinh M, Golib F, Liebermeister W, Poirier C, Tournier L, Klipp E, Fromion V, Goelzer A. Automated generation of bacterial resource allocation models. Metabolic Engineering. 2019 Sept 1;55:12–22.

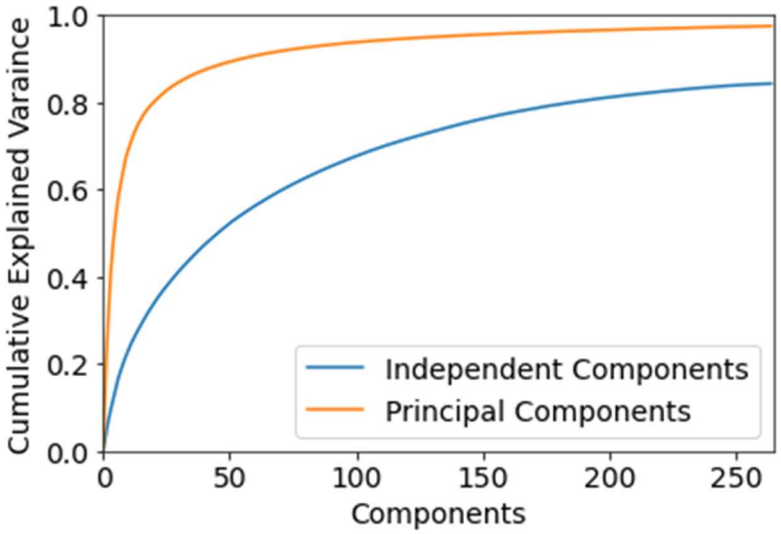
228. Salvy P, Hatzimanikatis V. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models. Nat Commun. 2020 Jan 13;11(1):30.

Appendix A. Proteome allocation is linked to transcriptional regulation through a modularized transcriptome



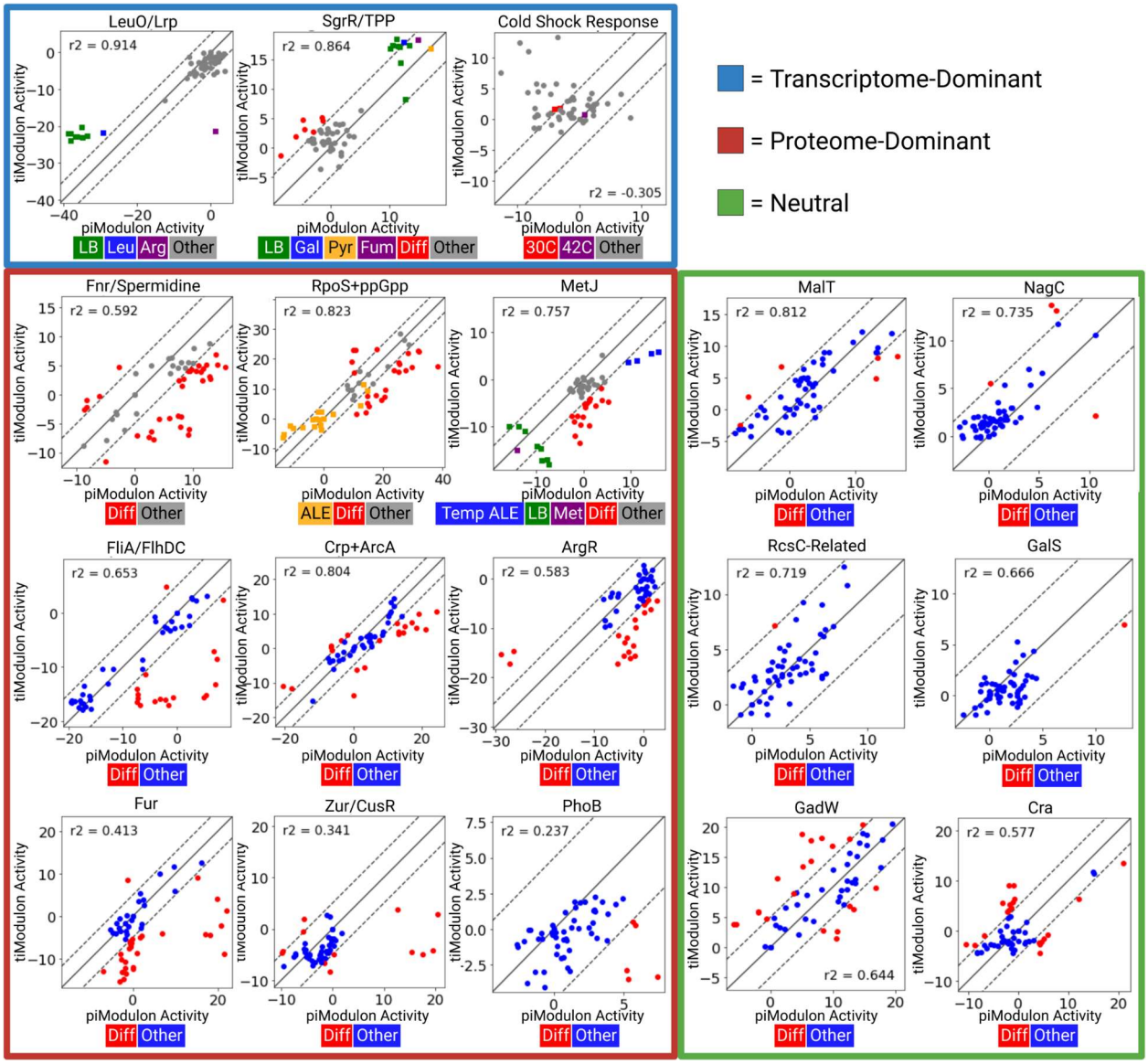
Supplemental Figure A.1: PRECISE1k replicate correlations**.**

R^2 values between biological replicates and random samples within PRECISE1k.



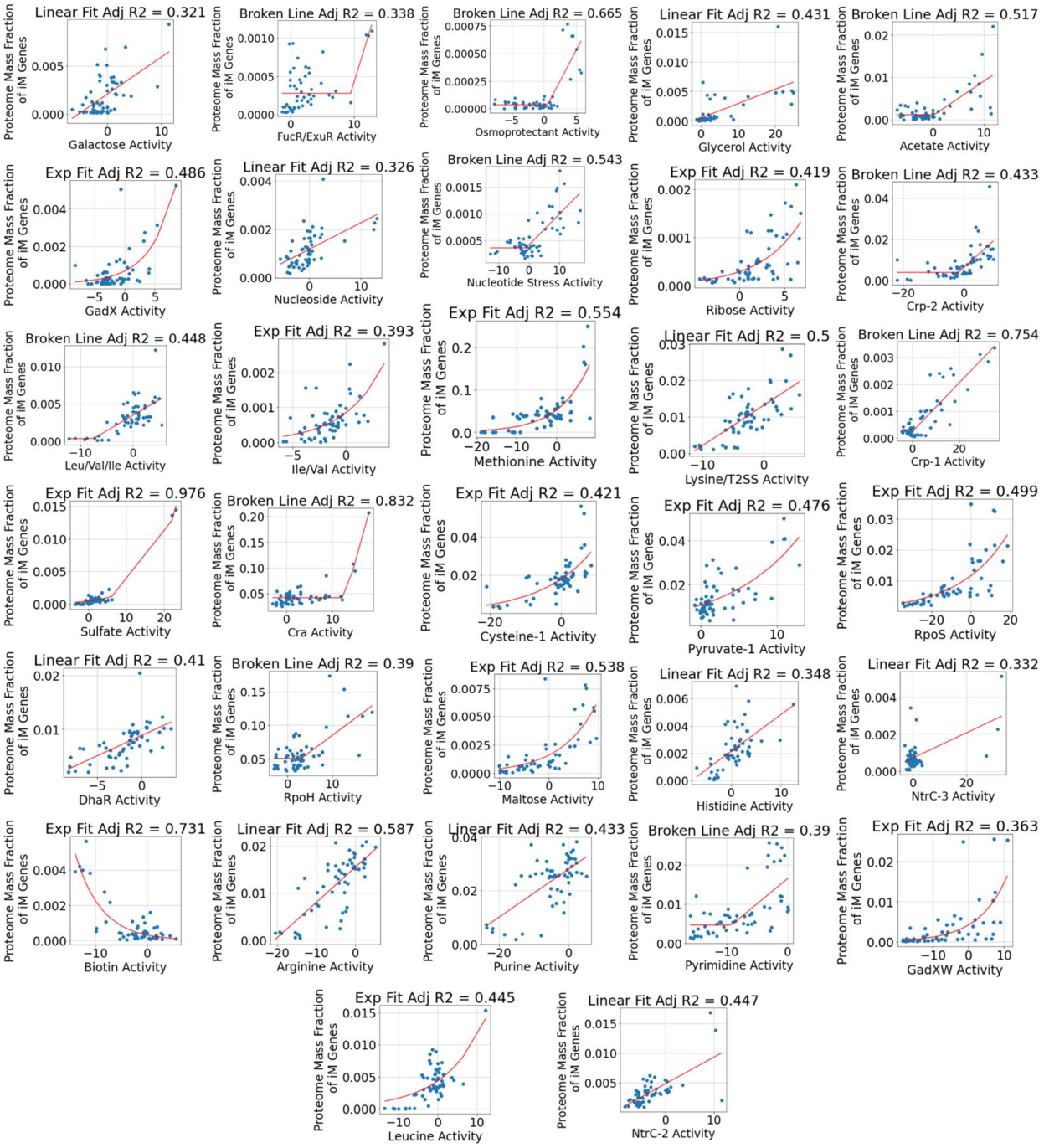
Supplemental Figure A.2: PRECISE1k explained variance.

Explained variance plot for PRECISE1k and its components.



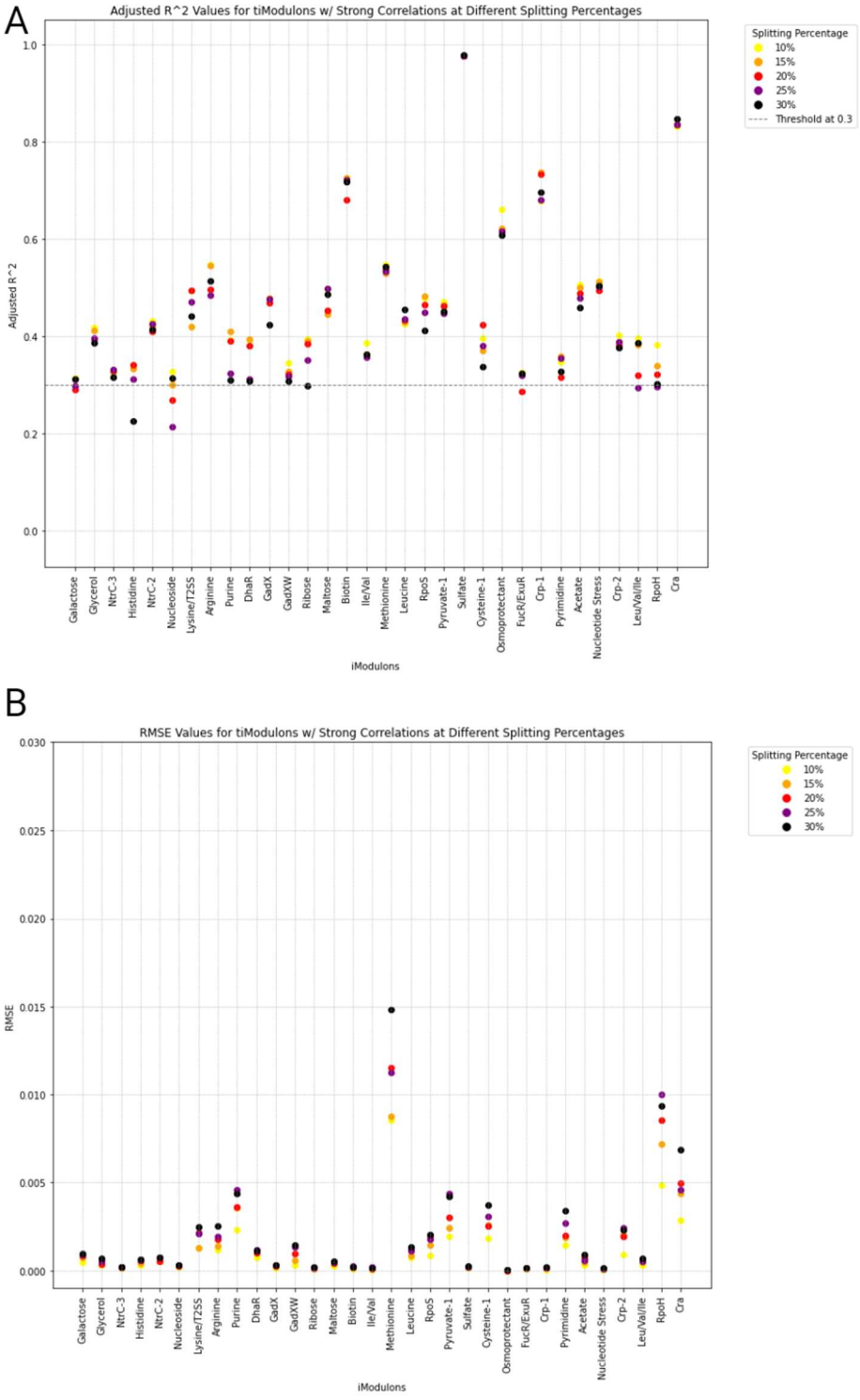
Supplemental Figure A.3: Comparing piModulon and tiModulon activities for all matched pairs.

Differential iModulon Activity (DiMA) plots for all matched iModulons between the two datasets. iModulons that are transcriptome-dominant (signal more active in the tiModulon than the piModulon) are highlighted in blue. iModulons that are proteome-dominant are highlighted in red, and iModulons that are neutral are highlighted in green. Activities are considered differentially activated for samples that lie outside the significance threshold (dashed line). Legends for each plot are placed below each plot. Abbreviations LB: Lysogeny Broth, Leu: Leucine Supplement, Arg: Arginine Supplement, Gal: Galactose Carbon Source, Pyr: Pyruvate Carbon Source, Fum: Fumerate Carbon Source, Diff: Differentially Activated, Temp: Temperature, Met: Methionine Supplement. n=57 conditions.



Supplemental Figure A.4: Scatter plots and regressions for all tiModulons with strong correlations.

Scatter plots for all tiModulon activities and their measured proteome mass fraction of the associated genes for tiModulons with strong correlations. tiModulons are charaterized based on which regression method (linear, exponential, and broken line) resulted in the best adjusted R^2 value. The regression method and R^2 value can be found above each plot, with the tiModulon name below. Individual samples are scatter plotted in blue, and the regression line is in red. n=57 conditions.

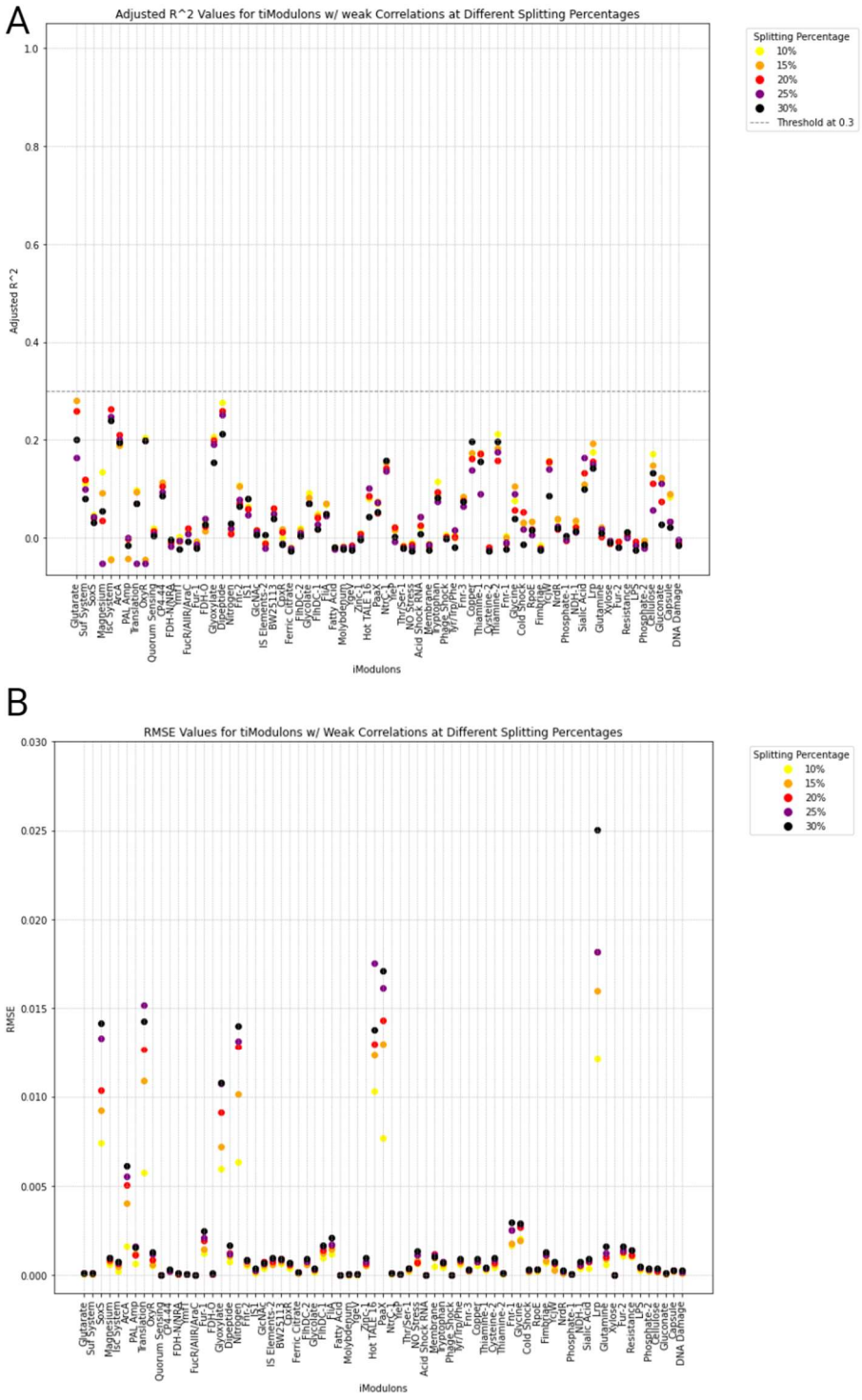


Supplemental Figure A.5: Leave-one-out cross-validation with additional holdouts for tiModulons w/ Strong Correlations.

(A) Adjusted R2 values for each regression model at various splitting percentages, only for tiModulons w/ strong correlations to proteome allocation. **(B)** Root Mean Square Errors for each regression model at various splitting percentages, only for tiModulons w/ strong correlations to proteome allocation.

Discussion of Supplemental Figure A.5:

The regression models for iModulons that are strongly correlated with their proteome allocation are actually fairly robust even for 57 samples. The Adjusted R2 values stay above the 0.3 threshold for almost all splitting percentages across iModulons except for a couple of extreme (>20%) splits (**Supplemental Figure A.5a**). This indicates that the models fit the data well. Furthermore, the root mean square errors (RMSEs) for the strongly correlated iModulons are also tight with an average standard deviation of 0.000438 (**Supplemental Figure A.5b**). This average standard deviation indicates that the model can accurately infer the proteome allocation for new data points. The scale of RMSE relates to the magnitude of proteome allocation, which is why the Cra, Methionine, and RpoH iModulons all have relatively high RMSEs (0.043367, 0.043133, and 0.033590 % of the proteome, respectively). These three iModulons are the three largest by proteome mass (see **Figure 2.5d** in main text) and thus vary the most, so seeing them stand out is expected.

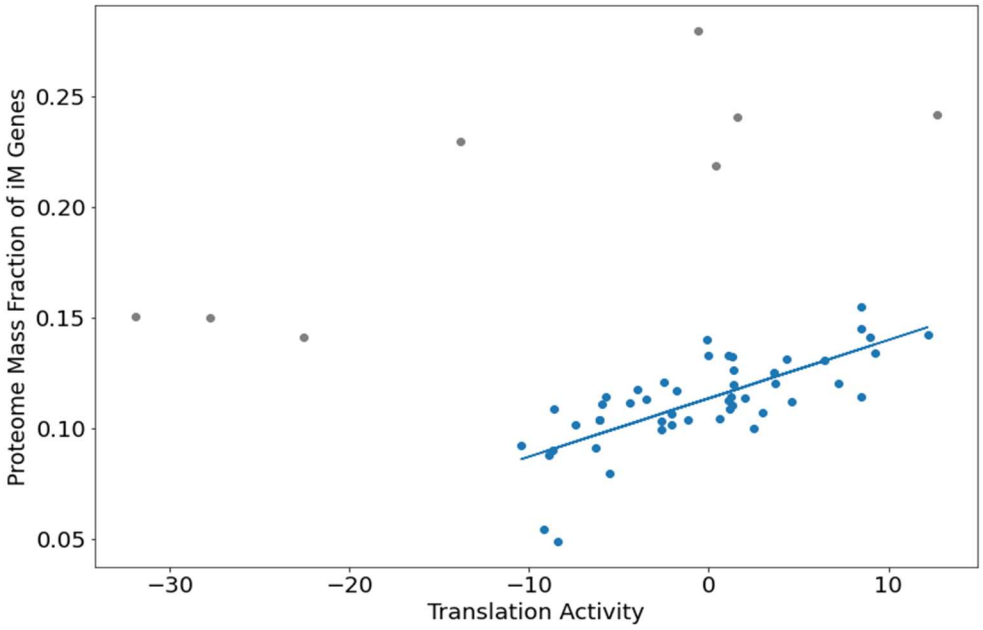


Supplemental Figure A.6: Leave-one-out cross-validation with additional holdouts for tiModulons w/ Weak Correlations.

(A) Adjusted R2 values for each regression model at various splitting percentages, only for tiModulons w/ weak correlations to proteome allocation. **(B)** Root Mean Square Errors for each regression model at various splitting percentages, only for tiModulons w/ weak correlations to proteome allocation.

Discussion of Supplemental Figure A.6:

The Adjusted R2 values consistently stay below 0.3, indicating that the models don't fit the data well (**Supplemental Figure A.6a**). The average standard deviation for the RMSEs for the weakly correlated iModulons is 0.000465, which is more than the average standard deviation for strongly correlated iModulons (**Supplemental Figure A.6b**). These iModulons are much smaller by proteome mass, so one would expect the standard deviation to be less since it should scale with the magnitude of proteome allocation (only Translation, Lrp, ArcA, and Glyoxylate are large by mass, but we unexpectedly also see other iModulons with high RMSEs, see **Figure 2.5d** in main text). These results indicate that not only do the models not fit the data well, they also don't predict new data well.



Supplemental Figure A.7: Proteome allocation of the translation tiModulon.

PRECISE1k Translation tiModulon activities plotted against their associated proteome mass fractions. There is a strong correlation between samples except for a few outliers (gray). $R^2 = 0.737$ without outliers, $R^2 = 0.135$ with outliers. n=57 conditions.

Supplemental Table A.1: Matched tiModulons and piModulons

piModulon	tiModulon	Correlation
MalT	Maltose	0.517757
RcsC-related	UC-4	0.503688
RcsC-related	pts ALE	0.505555
RpoS+ppGpp	RpoS	0.63325
NagC	GlcNAc	0.656724
FliA/FlhDC	FlhDC-2	0.336105
FliA/FlhDC	FliA	0.792489
Fur	Fur-1	0.68668
GalS	Galactose	0.535344
ArgR	Arginine	0.332504
Cold Shock Response	Cold Shock	0.489819
SgrR/Thiamine diphosphate	Thiamine-2	0.260406
SgrR/Thiamine diphosphate	Thiamine-1	0.65899
PhoB	Phosphate-1	0.427802
PhoB	tpiA KO	0.332344
Zur/CusR	Zinc-1	0.411696
Zur/CusR	Copper	0.2603
GadW	GadXW	0.581244
Translation-Null	Translation	0.391143
Fnr/Spermidine	Fnr-1	0.347905
Fnr/Spermidine	Fnr-3	0.401703
LeuO/Lrp	Lrp	0.65983
LeuO/Lrp	Pyruvate-1	0.264165
Crp+ArcA	Crp-2	0.532799
Crp+ArcA	Maltose	0.255092
Cra	Cra	0.680723
MetJ	Methionine	0.448365
Fimbriae	Fimbriae	0.471338

Appendix B. Laboratory evolution of synthetic electron transport system variants reveals a larger metabolic respiratory system and its plasticity

Supplemental Table B.1: Sample code used during running the ALE machine.

Strain	**ALE replicate number**	**Full ALE code**
*u*ETS-1H	Unevolved	Δ*nuoB*Δ*cyoB*_A0F0I1R
*e*ETS-1HA	ALE-1	Δ*nuoB*Δ*cyoB*_A9F63I1R1
*e*ETS-1HB	ALE-2	Δ*nuoB*Δ*cyoB*_A10F64I1R1
*e*ETS-1HC	ALE-3	Δ*nuoB*Δ*cyoB*_A11F64I1R1
*e*ETS-1HD	ALE-4	Δ*nuoB*Δ*cyoB*_A12F63I1R1
*u*ETS-2H	Unevolved	Δ*nuoB*Δ*cydB*Δ*appC*_A0F0I1R1
*e*ETS-2HA	ALE-1	Δ*nuoB*Δ*cydB*Δ*appC*_A13F59I1R1
*e*ETS-2HB	ALE-2	Δ*nuoB*Δ*cydB*Δ*appC*_A14F58I1R1
*e*ETS-2HC	ALE-3	Δ*nuoB*Δ*cydB*Δ*appC*_A15F57I1R1
*e*ETS-2HD	ALE-4	Δ*nuoB*Δ*cydB*Δ*appC*_A16F57I1R1
*u*ETS-3H	Unevolved	Δ*ndh*Δ*cyoB*_A0F0I1R1
*e*ETS-3HA	ALE-1	Δ*ndh*Δ*cyoB*_A5F59I1R1
*e*ETS-3HB	ALE-2	Δ*ndh*Δ*cyoB*_A6F58I1R1
*e*ETS-3HC	ALE-3	Δ*ndh*Δ*cyoB*_A7F56I1R1
*e*ETS-3HD	ALE-4	Δ*ndh*Δ*cyoB*_A8F59I1R1
*u*ETS-4H	Unevolved	Δ*ndh*Δ*cydB*Δa*ppC*_A0F0I1R1
*e*ETS-4HA	ALE-1	Δ*ndh*Δ*cydB*Δa*ppC*_A9F66I1R1
*e*ETS-4HB	ALE-2	Δ*ndh*Δ*cydB*Δa*ppC*_A10F71I1R1
*e*ETS-4HC	ALE-3	Δ*ndh*Δ*cydB*Δa*ppC*_A11F63I1R1
*e*ETS-4HD	ALE-4	Δ*ndh*Δ*cydB*Δa*ppC*_A12F60I1R1

Supplemental Table B.2: Key genetic changes in evolved ETS variants.

Strain	Media adaptation-related mutations	Condition-specific convergent gene mutation
*e*ETS-1HA	+5 bp between *hns* and *tdk* at genomic position 1293009, Δ82 bp between *pyrE* and *rph* at genomic position 3815859	None
*e*ETS-1HB	+5 bp between *hns* and *tdk* at genomic position 1293009, Δ82 bp between *pyrE* and *rph* at genomic position 3815859	None
*e*ETS-1HC	G to T substitution in *rpoC* at genomic position 4188513, Δ82 bp between *pyrE* and *rph* at genomic position 3815859	None
*e*ETS-1HD	+5 bp between *hns* and *tdk* at genomic position 1292997, Δ82 bp between *pyrE* and *rph* at genomic position 3815859	None
*e*ETS-2HA	A to T substitution in *rpoC* at genomic position 4187214	None
*e*ETS-2HB	C to T substitution in *rpoC* at genomic position 4185540	None
*e*ETS-2HC	A to G substitution in *rpoC* at genomic position 4187214	None
*e*ETS-2HD	C to T substitution in *rpoC* at genomic position 4185540	None
*e*ETS-3HA	Δ82 bp between *pyrE* and *rph* at genomic position 3815859	T to C substitution in *sdhA* at genomic position 756099
*e*ETS-3HB	Δ82 bp between *pyrE* and *rph* at genomic position 3815859	C to T substitution in *sdhA* at genomic position 756150
*e*ETS-3HC	Δ82 bp between *pyrE* and *rph* at genomic position 3815859	T to G substitution in *sdhA* at genomic position 756886
*e*ETS-3HD	Δ82 bp between *pyrE* and *rph* at genomic position 3815859	C to G substitution in *sdhA* at genomic position 756968
*e*ETS-4HA	C to A substitution in *rpoA* at genomic position 3440923	C to T substitution in *yjjX* at genomic position 4633743
*e*ETS-4HB	Δ3 bp in *rpoB* at genomic position 4183399	C to T substitution in *yjjX* at genomic position 4633634
*e*ETS-4HC	A to G substitution in *rpoC* at genomic position 4186578	C to T substitution in *yjjX* at genomic position 4633657
*e*ETS-4HD	C to T substitution in *rpoC* at genomic position 4189448	C to T substitution in *yjjX* at genomic position 4633657

The intergenic mutation between *pyrE* and *rph* was present in the respective unevolved strain itself and these are listed here for a better context of media adaptation. Other mutations present in unevolved strain have not been listed here. The genes mutated in all four evolved replicates of a strain are listed here as convergent gene mutations.

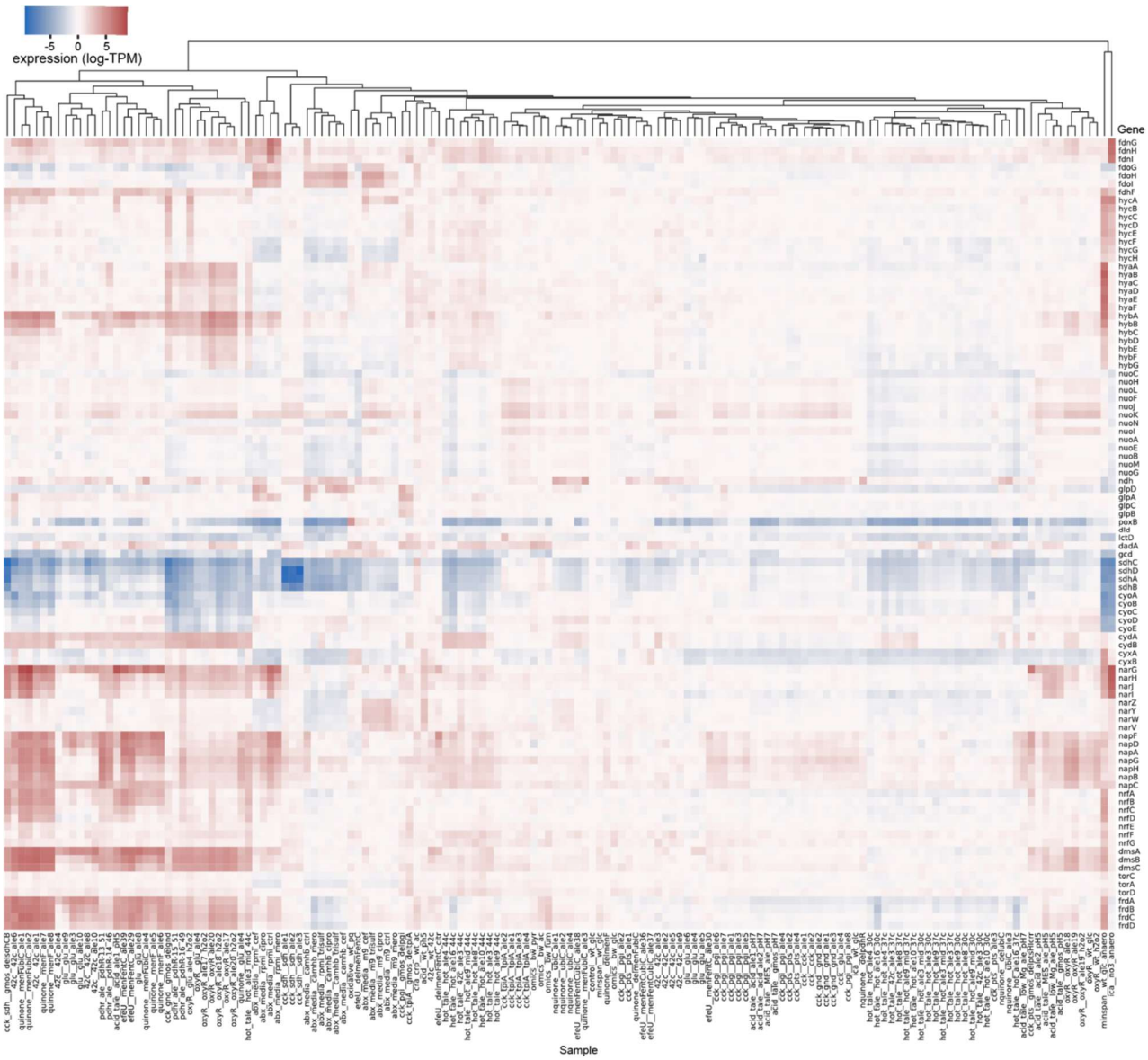
Supplemental Table B.3: Phenotypic characterization of the strains of the study.

Strain	Growth rate (h^{-1})	Glucose uptake rate (mmol/gDCW/h)	Acetate secretion rate (mmol/gDCW/h)
*u*ETS-1H	0.76, 0.79	12.77, 13.1	12.01, 12.35
*e*ETS-1HA	0.88, 0.9	13.57, 14.24	13.14, 13.61
*e*ETS-1HB	0.87, 0.82	13.61, 13.23	13.03, 12.88
*e*ETS-1HC	0.83, 0.82	13.84, 13.96	13.24, 12.93
*e*ETS-1HD	0.83, 0.83	13.35, 13.77	12.64, 13.06
*u*ETS-2H	0.65, 0.64	9.51, 9.26	6.41, 6.08
*e*ETS-2HA	0.87, 0.92	11.24, 11.71	7.76, 7.92
*e*ETS-2HB	0.9, 0.88	11.57, 11.67	7.77, 7.4
*e*ETS-2HC	0.9, 0.9	11.62, 11.68	7.36, 7.53
*e*ETS-2HD	0.89, 0.89	11.65, 11.76	7.98, 7.85
*u*ETS-3H	0.42, 0.43	7.68, 8.02	6.67, 6.51
*e*ETS-3HA	0.82, 0.83	12.6, 12.58	11.03, 11.26
*e*ETS-3HB	0.82, 0.86	12.35, 12.72	11.18, 11.26
*e*ETS-3HC	0.83, 0.84	12.78, 12.87	11.27, 10.97
*e*ETS-3HD	0.83, 0.84	11.92, 12.26	10.81, 11.07
*u*ETS-4H	0.53, 0.52	7.96, 8.18	3.95, 3.81
*e*ETS-4HA	0.85, 0.84	9.4, 9.01	4.24, 3.77
*e*ETS-4HB	0.86, 0.88	10.06, 10.47	5.25, 6.22
*e*ETS-4HC	0.79, 0.83	9.2, 9.3	6.14, 5.97
*e*ETS-4HD	0.8, 0.82	9.32, 9.49	4.96, 4.84

The values for two independent replicates are listed in the table. The levels of Succinate, Lactate, Formate, Ethanol, and Pyruvate were below the detection limit.

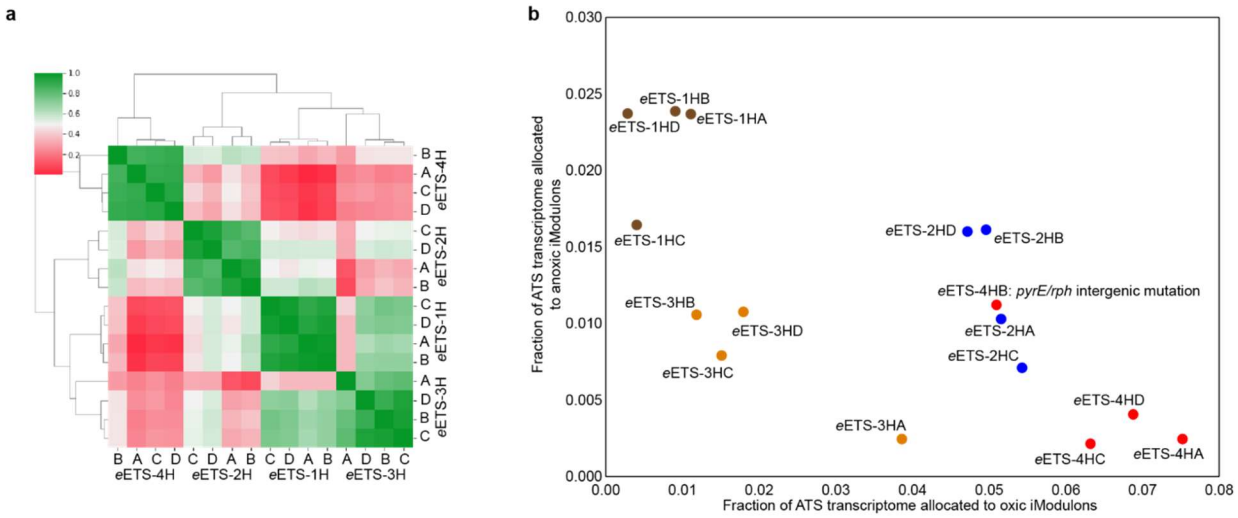
Supplemental Table B.4: Description of the ATS iModulons.

S. No.	iModulon	Genes (negative coefficients in red)
1	iModulon-13	*sdhB, cyoD, sdhD, cyoA, sdhA, sdhC, cyoC, fumA*
2	iModulon-b2287	*nuoB*
3	iModulon-8	*ndh, cyoB*
4	iModulon-11	*cydB, appB, cyoB, appC, ndh*
5	iModulon-b3366	*nirD*
6	iModulon-7	*nirB, grcA*
7	iModulon-2	*narI, narJ*
8	iModulon-1	*yliI, cyoD*
9	iModulon-16	*hyaB, hyaA, hyaC*
10	iModulon-9	*hybO, hybA*
11	iModulon-10	*tktB, adhP, fbaB, talA, poxB*



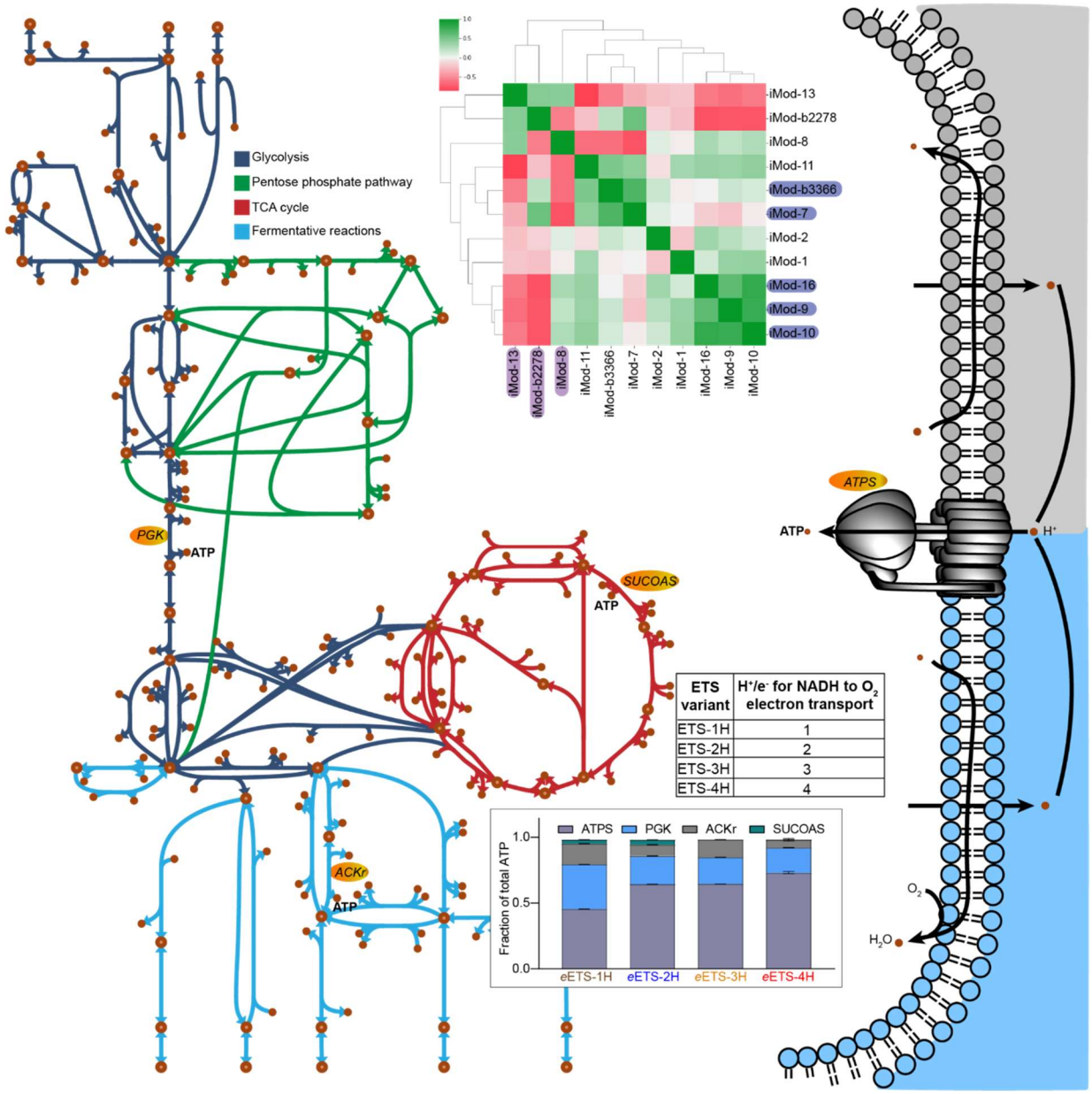
Supplemental Figure B.1: Expression levels of respiratory dehydrogenase and reductases of E. coli.

The entire RNA-seq compendium was explored and the conditions directly or indirectly associated with energy metabolism have been displayed.



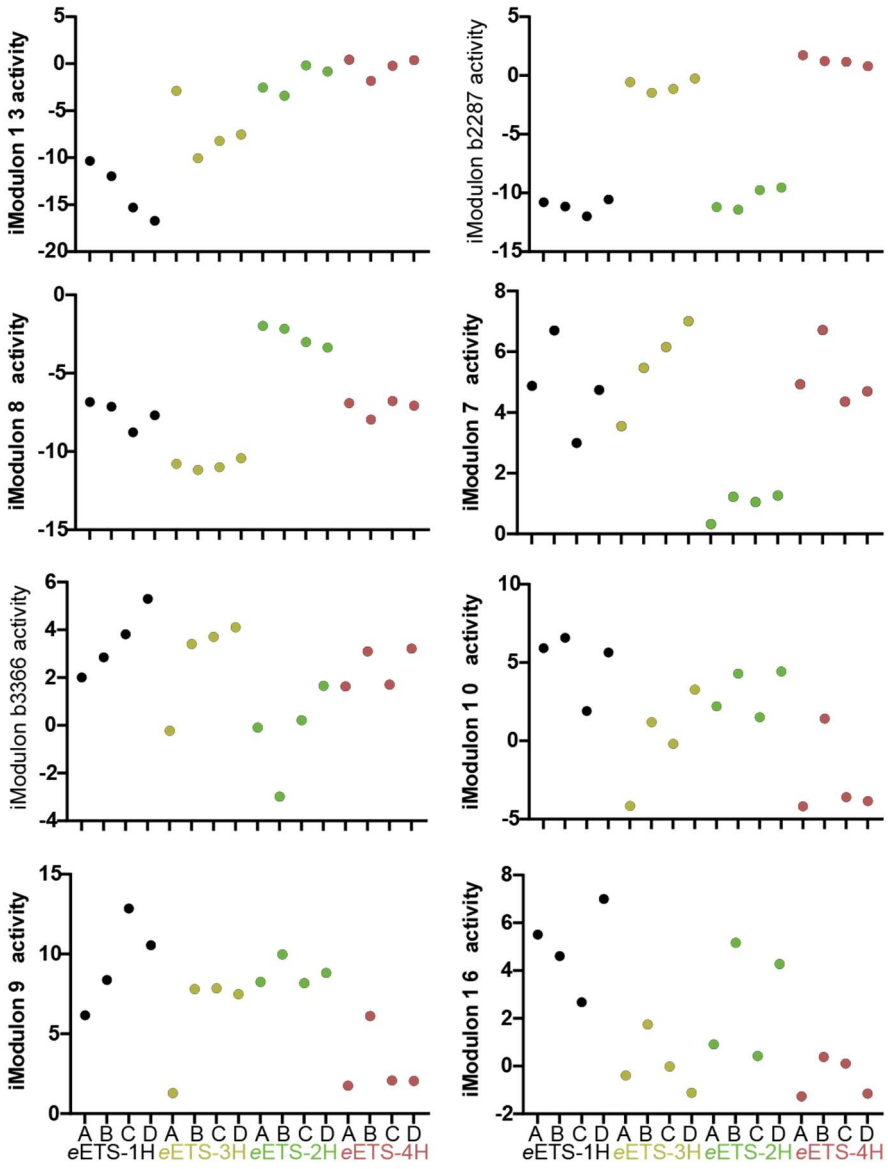
Supplemental Figure B.2: ATS iModulon correlations.

(a) Correlation among ETS variants based on the ATS iModulon activities. (b) The enlarged representation of figure 3d lists the replicate information and potential genetic change in the outlier replicate *e*ETS-4HB.



Supplemental Figure B.3: Scheme of Aero-type system (ATS).

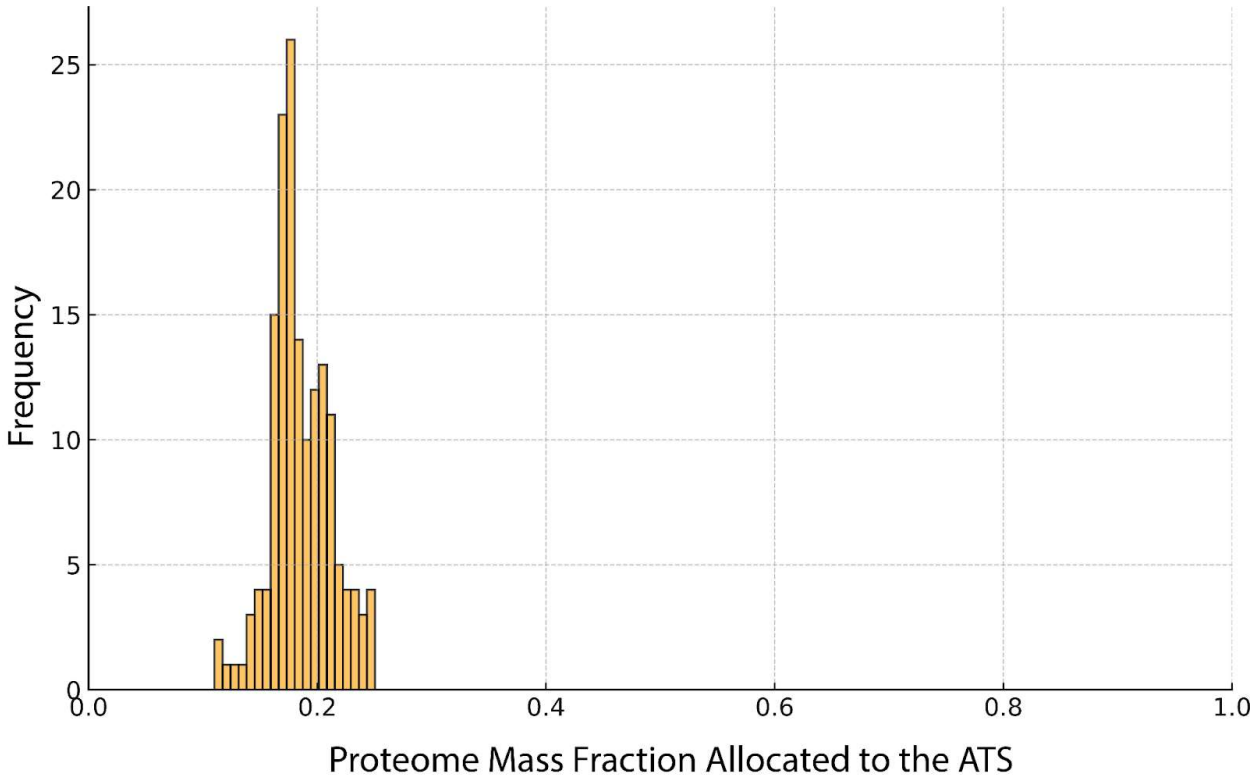
The pathways are showing reactions of ATS and major ATP production sites. Heatmap represents the correlation between the iModulons within ATS. iModulons showing a tradeoff are highlighted. The histogram from Figure 3 has been included to show adjustments within ATS to achieve a similar ATP yield. Part of the membrane highlighted in blue represents oxic ETS and another part highlighted in gray represents anoxic ETS. The list of ATS genes was generated based on COG and GO categories to include as many relevant genes as possible to represent pathways involved in ATP production, then filtered to remove genes that are never expressed in the multiple model simulations.



Supplemental Figure B.4: ATS iModulon Activities.

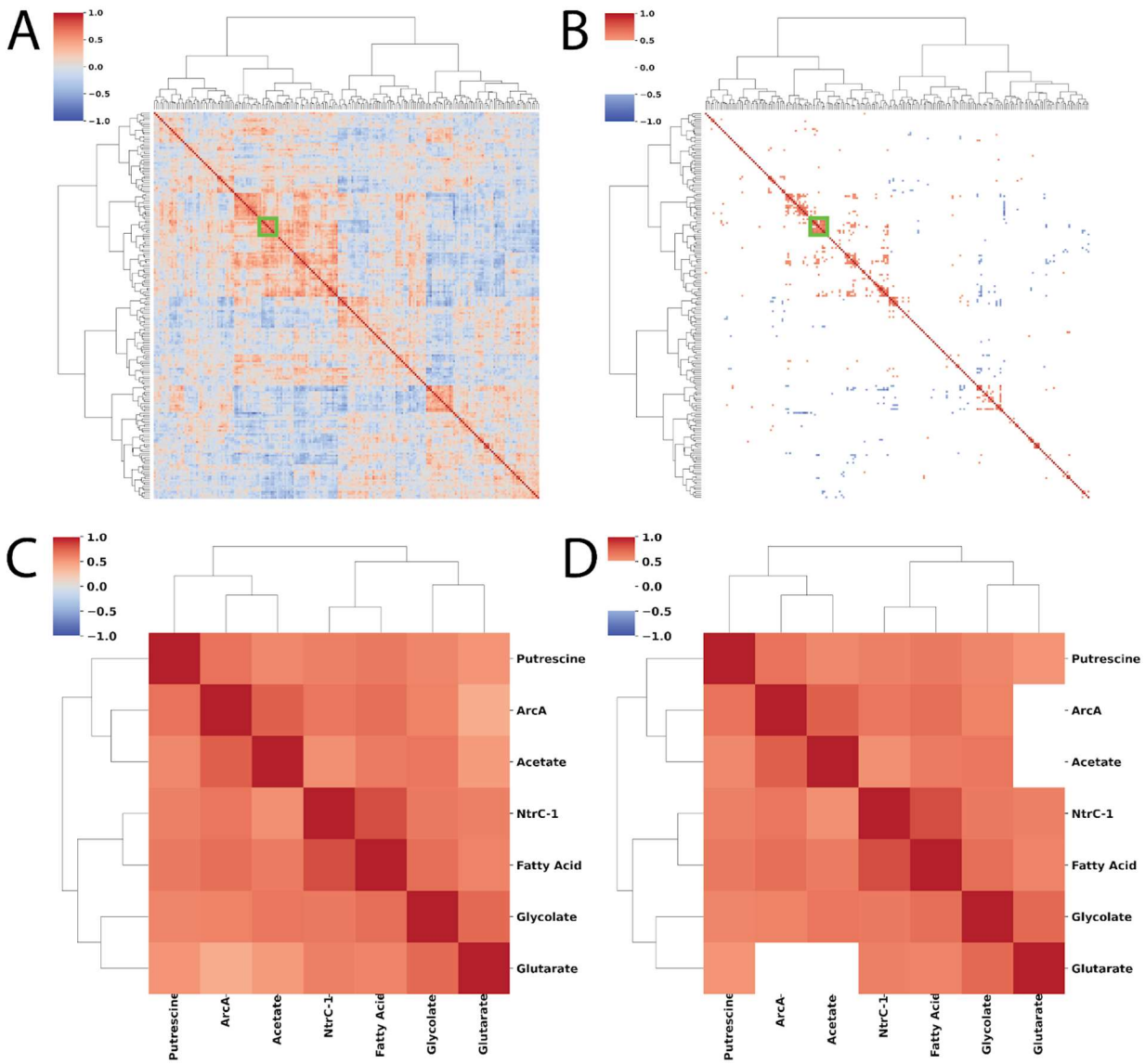
Activities of ATS iModulons in the unevolved and evolved ETS variants calculated using independent component analysis. The gene membership of corresponding iModulons is listed in Supplemental Table B.3.

Appendix C. Aerobicity stimulon in *Escherichia coli* revealed using multi-scale computational systems biology of adapted respiratory variants



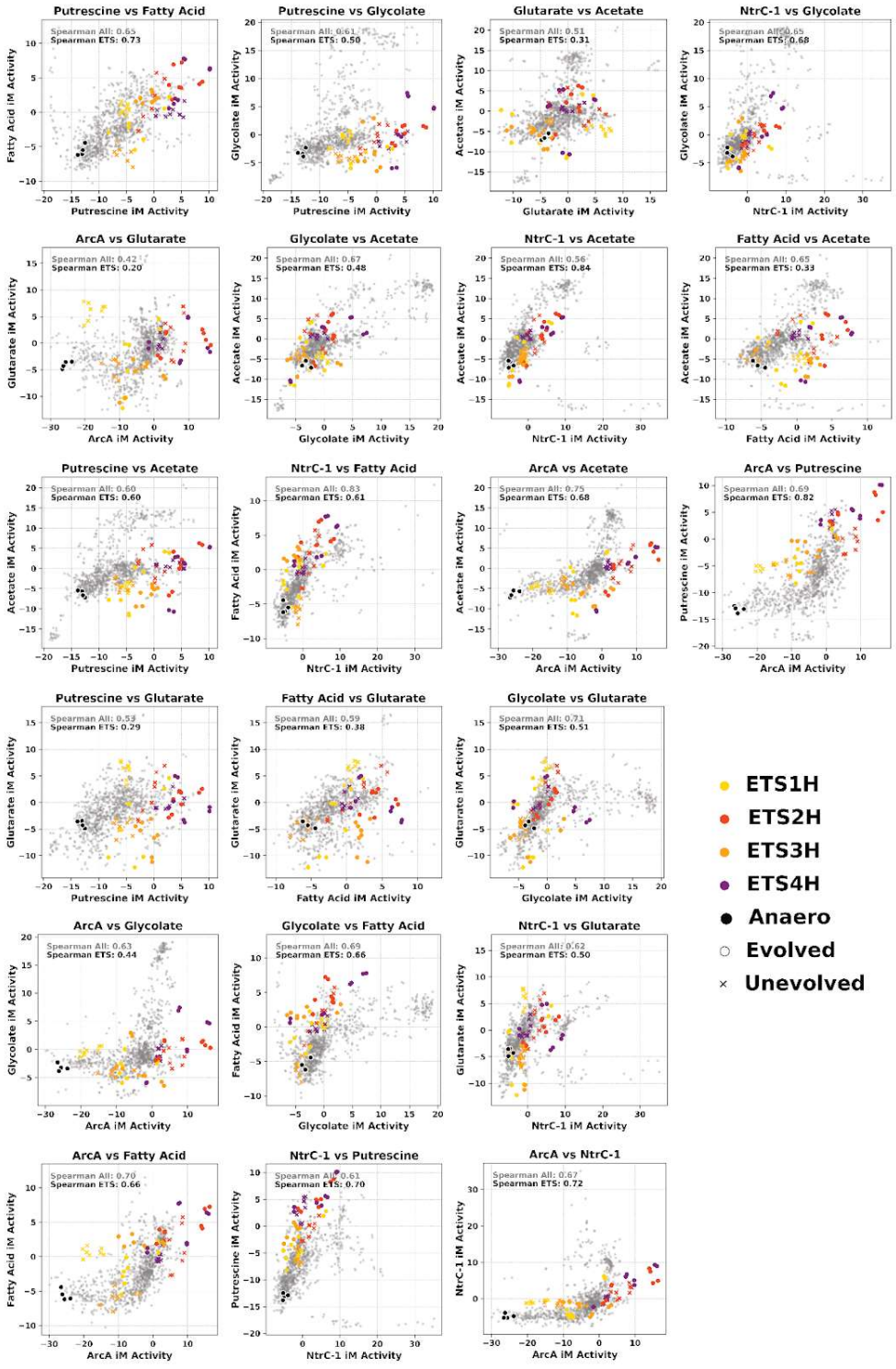
Supplemental Figure C.1: ATS proteome allocation across experimental conditions.

A histogram showing the proteome allocated to the Aero-Type System for all experimental samples from our in-house proteomic expression profiling compendium (n = 162 samples, mean = 0.19, std = 0.027).



Supplemental Figure C.2: PRECISE-1K activities correlation matrix.

A) Hierarchical clustering heatmap of the correlation matrix for all 201 iModulon activities across 1035 experimental conditions in our in-house transcriptomic compendium, PRECISE-1K. The green box highlights the iModulon cluster that describes aerobic shifts. Applying a correlation threshold of +/- 0.5 **(B)** results in only a few clusters standing out. **C)** A closer look at the seven iModulons in the green box (Putrescine, ArcA, Acetate, NtrC-1, Fatty Acid, Glycolate, Glutarate) that constitute the cluster with and without a threshold **(D)**.



Supplemental Figure C.3: All iModulon pairs in the aerobicity stimulon.

Scatter plots of iModulon activities for all samples (n=1035) in PRECISE-1K for all pairs of iModulons in the aerobicity stimulon. All glucose, succinate, and glycerol ETS variants are labeled. Anaerobic samples are labeled as well to highlight the aerobicity directionality.

Supplemental Table C.1: Sample code of unevolved strains used in this study.

Strain	Full ALE code
uETS-1H	Δ*nuoB*Δ*cyoB*_A0F0I1R1
uETS-2H	Δ*nuoB*Δ*cydB*Δ*appC*_A0F0I1R1
uETS-3H	Δ*ndh*Δ*cyoB*_A0F0I1R1
uETS-4H	Δ*ndh*Δ*cydB*Δ*appC*_A0F0I1R1

Supplemental Table C.2: Sample code of succinate evolved strains used in this study.

Strain	Full ALE code
SMOS-A	A2F102I1R1
SMOS-B	A3F101I1R1
SMOS-C	A4F102I1R1
eETS-1HA	Δ*nuoB*Δ*cyoB*_A7F48I1R1
eETS-1HB	Δ*nuoB*Δ*cyoB*_A5F51I1R1
eETS-1HC	Δ*nuoB*Δ*cyoB*_A6F52I1R1
eETS-2HA	Δ*nuoB*Δ*cydB*Δ*appC*_A9F98I1R1
eETS-2HB	Δ*nuoB*Δ*cydB*Δ*appC*_A11F99I1R1
eETS-2HC	Δ*nuoB*Δ*cydB*Δ*appC*_A10F100I1R1
eETS-3HA	Δ*ndh*Δ*cyoB*_A15F92I1R1
eETS-3HB	Δ*ndh*Δ*cyoB*_A13F94I1R1
eETS-3HC	Δ*ndh*Δ*cyoB*_A16F90I1R1
eETS-4HA	Δ*ndh*Δ*cydB*Δ*appC*_A20F97I1R1
eETS-4HB	Δ*ndh*Δ*cydB*Δ*appC*_A17F92I1R1
eETS-4HC	Δ*ndh*Δ*cydB*Δ*appC*_A18F97I1R1

Supplemental Table C.3: Sample code of glycerol evolved strains used in this study.

Strain	Full ALE code
GlyMOS-A	A2F72I1R1
GlyMOS-B	A1F73I1R1
GlyMOS-C	A3F79I1R1
eETS-1HA	Δ*nuoB*Δ*cyoB*_A8F35I1R1
eETS-1HB	Δ*nuoB*Δ*cyoB*_A6F35I1R1
eETS-1HC	Δ*nuoB*Δ*cyoB*_A7F35I1R1
eETS-2HA	Δ*nuoB*Δ*cydB*Δ*appC*_A12F35I1R1
eETS-2HB	Δ*nuoB*Δ*cydB*Δ*appC*_A9F35I1R1
eETS-2HC	Δ*nuoB*Δ*cydB*Δ*appC*_A11F34I1R1
eETS-3HA	Δ*ndh*Δ*cyoB*_A14F32I1R1
eETS-3HB	Δ*ndh*Δ*cyoB*_A13F33I1R1
eETS-3HC	Δ*ndh*Δ*cyoB*_A15F34I1R1
eETS-4HA	Δ*ndh*Δ*cydB*Δ*appC*_A20F65I1R1
eETS-4HB	Δ*ndh*Δ*cydB*Δ*appC*_A18F67I1R1
eETS-4HC	Δ*ndh*Δ*cydB*Δ*appC*_A19F70I1R1

Supplemental Table C.4: Key genetic changes in the succinate evolved ETS variants.

Strain	**Category 1** General resource optimization mutations	**Category 2** Electron transport system-related mutations	**Category 3.1** Succinate transport related mutations
SMOS-A	C to T substitution in *rpoC* at genomic position 4188572		+15 bp in *dcuB* at genomic position 4348677
SMOS-B	G to T substitution in *rpoC* at genomic position 4189508		+5 bp in *dcuB* at genomic position 4348595
SMOS-C	C to T substitution in *rpoC* at genomic position 4189448		*dctR*, *dctA* in the duplicated region at genomic position 3560455
eETS-1HA		C to T substitution in the intergenic region between *ycfP/ndh* at genomic position 1165960	*dctR*, *dctA* in the duplicated region at genomic position 3653201
eETS-1HB	*rpoD* in the duplicated region at genomic position 3131501 and *pyrE-rph* in the amplified region at genomic position 3805701	C to A substitution in the intergenic region between *ycfP/ndh* at genomic position 1165961	*dctR*, *dctA* in the duplicated region at genomic position 3584101 and *dcuD* in the amplified region at genomic position 3131501
eETS-1HC		C to T substitution in the intergenic region between *ycfP/ndh* at genomic position 1165960	*dctR*, *dctA* in the duplicated region at genomic position 3653201
eETS-2HA	*pyrE-rph* in the duplicated region at genomic position 3653201	G to T substitution in *pdhR* at genomic position 122248	*dctR*, *dctA* in the duplicated region at genomic position 3619401
eETS-2HB	G to T substitution in *rpoD* at genomic position 3214754	C to T substitution in *pdhR* at genomic position 122401	*dctR*, *dctA* in the duplicated region at genomic position 3619401
eETS-2HC	*rpoA* in the duplicated region at genomic position 3428501	C to T substitution in the intergenic region between *ycfP/ndh* at genomic position 1165961	*dctR*, *dctA* in the duplicated region at genomic position 3584201 and *dcuD* in the amplified region at genomic position 3366701
eETS-3HA	Δ1bp in *rpoC* at genomic position 4189434 and *pyrE-rph* in the duplicated region at genomic position 3805301		*dctR*, *dctA* in the duplicated region at genomic position 3653201
eETS-3HB	A to C substitution in *rpoC* at genomic position 4186605		*dctR*, *dctA* in the duplicated region at genomic position 3653201 and C to A substitution in *dcuB* at genomic position 4347763
eETS-3HC	T to G substitution in *rpoB* at genomic position 4183961		*dctR*, *dctA* in the duplicated region at genomic position 3652201
eETS-4HA	C to A substitution in *rpoC* at genomic position 4189438 and *pyrE-rph* in the duplicated region at genomic position 3721701		
eETS-4HB	Δ1 bp in *rpoC* at genomic position 4189451		
eETS-4HC	G to A substitution in *rpoB* at genomic position 4182880 and *pyrE-rph* in the duplicated region at the genomic position 3653201		*dctR*, *dctA* in the duplicated region at the genomic position 3653201

Supplemental Table C.5: Key genetic changes in the glycerol evolved ETS variants.

Strain	**Category 1** General resource optimization mutations	**Category 2** Electron transport system-related mutations	**Category 3.2** Glycerol retention-related mutations
GlyMOS-A	A to T substitution in *rpoB* at genomic position 4183102		T to A substitution in *glpK* at genomic position 4117005
GlyMOS-B	C to A and C to T substitution in *rpoC* at genomic positions 4188572 and 4188929		T to G substitution in *glpK* at genomic position 4116250
GlyMOS-C	G to A substitution in *rpoB* at genomic position 4182880		G to A substitution in *glpK* at genomic position 4117195
eETS-1HA	G to C substitution in *rpoC* at genomic position 4187603		G to C substitution in *glpK* at genomic position 4187603
eETS-1HB	C to T substitution in *rpoB* at genomic position 4182899		C to A substitution in *glpK* at genomic position 4117018
eETS-1HC	C to T substitution in *rpoC* at genomic position 4188572		C to T substitution in *glpK* at genomic position 4117060
eETS-2HA	C to A substitution in *rpoC* at genomic position 4187342		C to A substitution in *glpK* at genomic position 4116529
eETS-2HB	C to A substitution in *rpoC* at genomic position 4187342		C to A substitution in *glpK* at genomic position 4116529
eETS-2HC	A to G substitution in *rpoB* at genomic position 4182881		T to A substitution in *glpK* at genomic position 4117005
eETS-3HA	G to T substitution in *rpoB* at genomic position 4183597		T to G substitution in *glpK* at genomic position 4116250
eETS-3HB	Δ82 bp in *rph* at genomic position 3815859		G to T substitution in *glpK* at genomic position 4116658
eETS-3HC	Δ2 bp in *rph* at genomic position 3815882		G to A substitution in *glpK* at genomic position 4117195
eETS-4HA	C to A substitution in *rpoC* at genomic position 4188929		Single base substitution in the coding region of *glpK* at genomic position 4116532
eETS-4HB	*hns* and *tdk* in the amplified region at genomic position 1259601		C to A substitution in *glpK* at genomic position 4117201
eETS-4HC	A to C substitution in *rpoC* at genomic position 4188864		Single base substitution in the coding region of *glpK* at genomic position 4116532

Supplemental Table C.6: Phenotypic characterization of succinate evolved ETS variants

Strain	Growth rate (h^{-1})	Succinate uptake rate (mmol/gDCW/h)	Acetate secretion rate (mmol/gDCW/h)
WT	0.35, 0.40, 0.40	3.36, 4.14, 3.87	3.19, 2.68, 2.15
SMOS	0.57, 0.62, 0.53	3.71, 4.01, 3.49	ND, 0.55, ND
uETS-1H	0.08, 0.07, 0.07	0.92, 0.98, 0.93	ND, ND, ND
eETS-1H	0.53, 0.51, 0.38	3.36, 3.35, 4.52	1.40, 1.14, 2.17
uETS-2H	0.40, 0.40, 0.40	2.85, 3.04, 3.47	2.97, 3.11, 3.79
eETS-2H	0.57, 0.58, 0.55	2.96, 3.97, 4.87	ND, ND, ND
uETS-3H	0.35, 0.34, 0.32	2.00, 2.76, 3.16	1.46, 2.31, 2.16
eETS-3H	0.53, 0.58, 0.61	6.41, 5.02, 5.35	2.75, 3.81, 3.73
uETS-4H	0.45, 0.39, 0.41	2.13, 2.56, 2.40	0.98, 1.33, 1.33
eETS-4H	0.60, 0.67, 0.62	4.42, 3.47, 3.88	3.01, 2.72, 2.60

For phenotypic characterization, lineage A has been used. Values for independent biological replicates have been shown in the table. ND: not detected.
Note - uETS variants have been grown in succinate minimal media and used for comparison.

Supplemental Table C.7: Phenotypic characterization of glycerol evolved ETS variants.

Strain	Growth rate (h^{-1})	Glycerol uptake rate (mmol/gDCW/h)	Acetate secretion rate (mmol/gDCW/h)
WT	0.45, 0.42, 0.43	13.14, 10.63, 10.07	0.23, 0.81, 0.26
GlyMOS	0.67, 0.70, 0.75	10.21, 10.90, 14.22	3.46, 3.48, 4.08
uETS-1H	0.39, 0.36, 0.43	8.28, 8.56, 10.48	1.99, 1.99, 4.88
eETS-1H	0.82, 0.69, 0.65	14.44, 13.85, 10.12	7.00, 6.59, 4.79
uETS-2H	0.43, 0.43, 0.46	6.98, 6.57, 6.69	0.34, 0.31, ND
eETS-2H	0.72, 0.68, 0.71	22.91, 16.28, 13.46	3.02, 3.65, 3.20
uETS-3H	0.43, 0.43, 0.43	10.34, 10.55, 10.33	2.53, 2.29, 2.13
eETS-3H	0.63, 0.62, 0.60	16.91, 14.31, 13.82	7.09, 6.27, 5.90
uETS-4H	0.51, 0.50, 0.48	6.77, 6.19, 5.92	ND, ND, ND
eETS-4H	0.58, 0.64, 0.55	11.41, 10.69, 10.08	2.86, 2.62, 2.44

For phenotypic characterization, lineage A has been used. Values for independent biological replicates have been shown in the table. ND: not detected.
Note - uETS variants have been grown in glycerol minimal media and used for comparison.

Appendix D. COBRAme.org: A web application for solving and constraining metabolism and expression models

Supplemental Table D.1: K_{eff} Modifications for DynamicME Replication

Reaction	New K_{eff}
GLYCtpp_REV_SPONT	0.0001
GLYK_FWD_GLYCEROL-KIN-CPLX_mod_mg2	25.7013
GLYK_FWD_GLYCEROL-KIN-CPLX_mod_mn2	25.7013
ICDHyr_FWD_ISOC ITHASE-CPLX_mod_mn2	198.333
ICDHyr_FWD_ISOCITHASE-CPLX_mod_mg2	198.333
CYTBO3_4pp_FWD_CYT-O-UBIOX-CPLX_mod_pheme_mod_hemeO_mod_cu2	300
CYTBD2pp_FWD_APP-UBIOX-CPLX_mod_pheme_mod_hemed	33.3358
CYTBDpp_FWD_APP-UBIOX-CPLX_mod_pheme_mod_hemed	33.3358
CYTBDpp_FWD_CYT-D-UBIOX-CPLX_mod_pheme_mod_hemed	33.4129