أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Developing Datasets and Algorithms for Metabolic Modelling and Engineering

Hussain Alharbi, Lya Chiñas, Radhakrishnan Mahadevan

hussain.alharbi@kaust.edu.sa, lya.chinas@mail.utoronto.ca, krishna.mahadevan@utoronto.ca

Abstract

Microbial growth prediction from genomic and environmental information remains a fundamental challenge limiting the development of predictive microbiology and its applications in biomanufacturing, bioremediation, and microbiome engineering. A major barrier is the lack of standardized, reproducible, and machine-learning-ready datasets that integrate genotype, phenotype, and environmental context. This study presents an open-source framework for generating and modeling microbial growth data through standardized experimental characterization and genomic feature engineering. Twenty-five microbial reference strains spanning four phyla were evaluated across 11 cultivation media under aerobic and anaerobic conditions and two temperature regimes. Growth dynamics were quantified using optical density measurements, with kinetic parameters including maximum growth rate, maximum optical density, lag phase duration, and area under the growth curve extracted through sigmoid curve fitting. Experimental reproducibility was assessed across two independent laboratory sites. Genome-derived features were generated through protein prediction, functional annotation, KEGG Orthology assignment, and metabolic pathway reconstruction, enabling integration of biological features with environmental variables for machine learning-based growth prediction. The resulting dataset and computational framework are expected to establish a scalable open-source foundation for genotype-to-phenotype modeling, supporting improved microbial cultivation strategies, metabolic engineering, and data-driven biotechnology development.

Date: July 25, 2026

Contents

1 Introduction **2**

1.1 Problem Motivation and Real-World Impact . . . 2
1.2 Problem Statement . . . 2
1.3 Assumptions and Limitations . . . 2
1.4 Proposed System Overview . . . 3

2 Related Work **3**

2.1 Genome-Scale Metabolic Modeling . . . 3
2.2 Microbial Phenotype Prediction and Machine Learning Approaches . . . 4
2.3 Large-Scale Biological Data Generation and Standardization . . . 4
2.4 Limitations of Existing Approaches . . . 4
2.5 Positioning of the Proposed Framework . . . 5

3 Methodology **5**

3.1 System Overview . . . 5
3.2 Experimental Data Generation and Standardization . . . 6
3.3 Data Processing and Quality Control . . . 6
3.4 Genome-Based Feature Extraction . . . 6
3.5 Feature Integration and Machine Learning Model . . . 7

4 Evaluation Procedure **7**

4.1 Computational Tools and Frameworks . . . 7
4.2 Dataset Construction . . . 7
4.3 Performance Evaluation Metrics . . . 8

5 Results **8**

5.1 Microbial Growth Characterization Across the Pilot Panel . . . 8
5.2 Oxygen Availability Is the Dominant Determinant of Growth . . . 8
5.3 Media Composition Shapes Biomass Yield and Growth Rate Differently . . . 9
5.4 Growth Rate and Biomass Yield Are Weakly Correlated and Should Be Treated as Distinct Prediction Targets . . . 9
5.5 Experimental Reproducibility Across Sites . . . 10
5.6 Data Quality Control . . . 10
5.7 Strain Metadata Standardization Pipeline . . . 10
5.8 Genome Annotation Pipeline: Installation and Preliminary Testing . . . 10
5.9 A Modular, License-Independent Alternative Pipeline . . . 10
5.10 Planned Benchmark of Annotation Pipelines . . . 11
5.11 Future Evaluation and Benchmarking of Predictive Models . . . 11

6 Discussion **11**

7 Conclusion **12**

1 Introduction

Microorganisms are fundamental to many biological and industrial processes, including biopharmaceutical production, sustainable chemical manufacturing, agriculture, environmental remediation, and microbiome engineering [**Nielsen2016**, **Lee2016**]. The ability to predict microbial behavior under specific genetic and environmental conditions would significantly accelerate the design, optimization, and deployment of microbial systems. However, despite the rapid expansion of microbial genome sequencing technologies, translating genomic information into predictable phenotypic outcomes remains a major challenge due to the complex relationship between genotype, metabolism, and environmental factors [**Zampieri2019**].

1.1 Problem Motivation and Real-World Impact

Modern sequencing technologies have generated extensive repositories of microbial genomes, providing unprecedented access to genetic information. However, genomic information alone does not provide a direct understanding of microbial function or performance under different cultivation conditions. Microbial growth is governed by complex interactions between metabolic networks, regulatory mechanisms, nutrient availability, and environmental conditions, making phenotype prediction difficult using sequence information alone [**Medema2022**].

A reliable genotype-to-phenotype prediction framework could transform microbial engineering by reducing the need for extensive experimental screening. Such predictive capabilities could enable faster identification of optimal microbial hosts for industrial fermentation, improve metabolic engineering strategies, and support the design of microbial communities for environmental and medical applications. However, current approaches are limited by the absence of large-scale, standardized, and experimentally validated datasets that connect microbial genetic information with quantitative phenotypic measurements.

1.2 Problem Statement

The central problem addressed in this research is the inability to accurately predict microbial growth behavior from genomic and environmental information due to the lack of reproducible experimental datasets and integrated computational frameworks capable of connecting genotype, metabolism, and phenotype.

Existing microbial databases primarily provide genomic sequences, annotations, or isolated experimental measurements, while large-scale phenotype datasets generated under standardized conditions remain limited. Furthermore, experimental variability between laboratories introduces uncertainty that can restrict the generalizability of machine learning models. Therefore, there is a need for a framework that combines controlled experimental characterization, independent validation, automated biological feature extraction, and predictive modeling.

This research addresses the following question:

Can standardized and reproducible microbial growth measurements, combined with genome-derived metabolic features and machine learning approaches, enable scalable prediction of microbial growth across diverse strains and environmental conditions?

1.3 Assumptions and Limitations

The proposed framework assumes that microbial growth characteristics measured under controlled laboratory conditions provide meaningful representations of strain-specific phenotypic behavior. Additionally, it assumes that genomic features derived from functional annotation and metabolic pathway reconstruction capture biologically relevant information associated with microbial

metabolic capacity.

However, microbial phenotypes are influenced by complex biological mechanisms, including gene regulation, epigenetic effects, evolutionary adaptation, and microbial interactions, which may not be fully represented by genome-derived features alone. Furthermore, laboratory conditions represent simplified environments and may not capture all factors influencing microbial behavior in natural ecosystems.

1.4 Proposed System Overview

To address these challenges, this work develops an integrated experimental-computational framework for predictive microbial growth modeling. The proposed system combines standardized wet-laboratory characterization with computational biology and machine learning pipelines to establish a scalable genotype-to-phenotype prediction platform.

The experimental component consists of systematic microbial growth characterization across controlled environmental conditions. Multiple independent wet-laboratory sites are used to validate reproducibility and reduce experimental bias by generating comparable growth measurements across locations. The resulting growth profiles provide quantitative phenotypic data, including kinetic parameters such as maximum growth rate, maximum biomass accumulation, lag phase duration, and growth curve characteristics.

The computational framework processes experimental and genomic information through an automated pipeline. Raw experimental measurements are cleaned, standardized, and structured into machine-learning-ready datasets. Genome sequences are analyzed through genomic metadata retrieval, protein prediction, functional annotation, metabolic pathway identification, and biological feature extraction. These metabolic and genomic representations are integrated with experimentally measured growth phenotypes and environmental parameters to train machine learning models capable of predicting microbial growth under previously characterized conditions.

The long-term objective of this framework is to generate a large-scale microbial growth atlas consisting of approximately one million experimentally validated datasets derived from 1,000 microbial strains evaluated across 1,000 environmental conditions. By integrating reproducible experimental measurements with computational modeling, this work aims to establish a foundation for scalable predictive microbiology and accelerate microbial engineering applications.

2 Related Work

2.1 Genome-Scale Metabolic Modeling

Genome-scale metabolic models (GEMs) have been widely used to represent microbial metabolism and predict cellular behavior based on reconstructed metabolic networks. Constraint-based modeling approaches, particularly flux balance analysis (FBA), have enabled the prediction of metabolic flux distributions under specific environmental conditions and have become important tools in systems biology and metabolic engineering [**Orth2010FBA**]. Large-scale platforms such as BiGG Models have further expanded the accessibility of curated metabolic reconstructions for numerous organisms [**King2016BiGG**].

However, traditional genome-scale metabolic modeling approaches rely heavily on manually curated metabolic networks and predefined biological constraints. Although these models provide mechanistic interpretations of cellular metabolism, their predictive accuracy is often limited by incomplete annotations, inaccurate assumptions regarding cellular regulation, and insufficient experimental validation across diverse environmental conditions [**Zampieri2019**]. Therefore,

while GEMs provide valuable biological representations, they do not independently solve the challenge of predicting microbial phenotypes at scale.

2.2 Microbial Phenotype Prediction and Machine Learning Approaches

Recent advances in machine learning have introduced new approaches for predicting microbial phenotypes from genomic and environmental information. Machine learning models have demonstrated potential in identifying complex nonlinear relationships between genetic features and microbial traits that are difficult to capture using traditional mechanistic models [**Mahieu2021**]. These approaches have been applied to tasks including metabolic engineering, microbial classification, and prediction of antimicrobial resistance.

Despite these advances, many existing machine learning approaches are constrained by limited training datasets and inconsistent phenotype measurements. Microbial phenotype data are often collected from heterogeneous sources with varying experimental protocols, cultivation conditions, and measurement methods. This lack of standardization reduces model generalizability and makes it difficult to determine whether predictive performance reflects biological relationships or experimental bias [**Medema2022**].

2.3 Large-Scale Biological Data Generation and Standardization

Large biological databases have transformed research by providing standardized repositories of genomic and molecular information. Resources such as UniProt and KEGG provide extensive protein annotations, functional classifications, and pathway information that enable computational analysis of microbial systems [**UniProt2023**, **Kanehisa2023**]. However, these resources primarily describe genomic and biochemical information rather than quantitative phenotypic responses under controlled environmental conditions.

Several large-scale experimental initiatives have attempted to address the challenge of generating standardized biological datasets. Similar to benchmark datasets developed in other machine learning domains, high-quality biological datasets require consistent data collection, metadata standardization, and experimental reproducibility. However, microbial growth datasets remain limited in scale and frequently lack independent experimental validation across laboratories.

2.4 Limitations of Existing Approaches

Existing approaches have advanced individual components of microbial prediction, including metabolic reconstruction, genomic annotation, and machine learning-based phenotype prediction. However, several limitations remain:

- **Limited experimental scalability:** Most studies characterize a relatively small number of microbial strains or environmental conditions, restricting the diversity of available training data.
- **Insufficient experimental reproducibility:** Many phenotype datasets are generated from single laboratories without independent validation, making it difficult to assess robustness across experimental environments.
- **Limited integration of biological knowledge:** Machine learning approaches often rely on raw genomic features without incorporating biologically meaningful representations such as metabolic pathways and functional annotations.
- **Fragmented computational workflows:** Existing tools typically address individual tasks, such as genome annotation, pathway reconstruction, or model prediction, rather than providing an integrated end-to-end framework.

2.5 Positioning of the Proposed Framework

This work addresses these limitations by developing an integrated experimental-computational framework that combines standardized microbial growth characterization, independent wet-laboratory validation, genomic metadata retrieval, metabolic pathway reconstruction, and machine learning-based growth prediction.

Unlike previous approaches that primarily rely on existing heterogeneous datasets or computationally reconstructed phenotypes, the proposed framework generates experimentally validated phenotype data under controlled conditions and directly links these measurements with genome-derived biological features. By establishing a scalable pipeline capable of generating and analyzing large numbers of strain-condition combinations, this work aims to enable more robust genotype-environment-phenotype prediction and advance the development of predictive microbiology platforms.

3 Methodology

3.1 System Overview

This work proposes an integrated open-source experimental-computational framework for predicting microbial growth from genomic and environmental information. The system combines standardized wet-laboratory phenotyping, computational genome analysis, metabolic pathway reconstruction, and machine learning-based predictive modeling.

The overall workflow consists of five major stages: (1) standardized microbial growth characterization, (2) experimental quality control and data standardization, (3) genomic metadata retrieval and metabolic feature extraction, (4) integration of biological and environmental features for machine learning model development, and (5) evaluation of model performance for microbial growth prediction.

The experimental component generates quantitative growth phenotypes by characterizing microbial strains under controlled cultivation conditions. To improve reproducibility and reduce experimental bias, growth experiments are performed across two independent wet-laboratory sites, enabling cross-validation of generated phenotypic measurements. The computational component transforms raw experimental and genomic information into machine-learning-ready representations by extracting biologically meaningful features, including metabolic pathways, functional annotations, and environmental variables. The complete system architecture is illustrated in Figure 1.

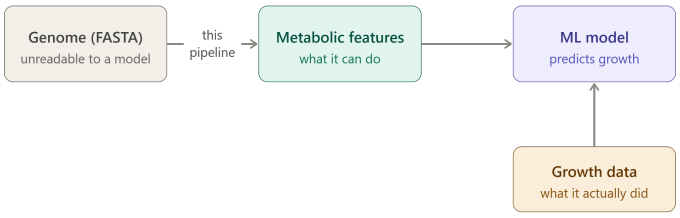


Figure 1 Overview of the proposed experimental-computational framework for microbial growth prediction. The framework integrates standardized wet-lab phenotyping, genome-based feature extraction, metabolic pathway reconstruction, and machine learning prediction.

3.2 Experimental Data Generation and Standardization

The first component of the framework focuses on generating standardized microbial growth datasets through controlled laboratory experiments. Microbial reference strains are cultivated under defined environmental conditions, including different growth media, oxygen availability states, and temperature regimes. Growth dynamics are measured through optical density (OD) measurements collected over time.

The resulting growth curves are processed to extract quantitative kinetic parameters describing microbial growth behavior. These parameters include maximum growth rate (μ_{max}), maximum optical density (OD_{max}), lag phase duration (λ), and area under the growth curve (AUC). Growth curve fitting is performed using sigmoid-based models:

$$OD(t) = f(t; \theta) \tag{1}$$

where $OD(t)$ represents microbial biomass accumulation over time and θ represents the fitted growth parameters.

To evaluate experimental reproducibility, identical experimental conditions are characterized across two independent wet-laboratory sites. Agreement between sites is assessed through comparison of extracted growth parameters and statistical evaluation of measurement variability.

3.3 Data Processing and Quality Control

Raw experimental measurements undergo preprocessing to remove inconsistencies, identify experimental errors, and standardize metadata representation. The preprocessing pipeline includes:

- Removal of invalid or incomplete measurements.
- Normalization of growth measurements across experiments.
- Standardization of strain identifiers and experimental metadata.
- Extraction of growth curve parameters.
- Integration of environmental conditions with phenotype measurements.

The processed dataset is structured into a standardized representation:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \tag{2}$$

where x_i represents the biological and environmental features associated with experimental condition i, and y_i represents the measured microbial growth phenotype.

3.4 Genome-Based Feature Extraction

To incorporate biological information into predictive models, genome sequences associated with each microbial strain are processed through an automated computational pipeline. Genome metadata are retrieved and analyzed using bioinformatics tools for sequence characterization, protein prediction, and functional annotation.

The pipeline includes:

1. Genome metadata retrieval from public repositories.
2. Protein sequence prediction and functional annotation.
3. Assignment of functional categories using databases such as KEGG Orthology.
4. Reconstruction of metabolic pathways.
5. Generation of pathway-based and genomic feature representations.

The resulting genomic representation is defined as:

$$G_i = [g_1, g_2, ..., g_n] \tag{3}$$

where G_i represents the metabolic and genomic feature vector associated with microbial strain i.

3.5 Feature Integration and Machine Learning Model

The predictive component combines genomic features, metabolic pathway information, and environmental parameters to predict microbial growth behavior. The integrated feature representation is expressed as:

$$X_i = [G_i, E_i, M_i] \tag{4}$$

where G_i represents genome-derived features, E_i represents environmental variables, and M_i represents metabolic pathway features.

The objective of the predictive model is to learn a function:

$$f(X_i) \rightarrow y_i \tag{5}$$

where f represents the machine learning model and y_i represents the expected microbial growth phenotype.

Different machine learning approaches can be evaluated, including regression-based models, ensemble learning methods, and neural network architectures. Dimensionality reduction techniques are applied when necessary to improve feature representation, reduce computational complexity, and minimize feature redundancy.

4 Evaluation Procedure

4.1 Computational Tools and Frameworks

The proposed framework integrates multiple computational tools for biological data processing and machine learning development. The computational environment includes:

- Python-based data processing and machine learning libraries, including NumPy, Pandas, Scikit-learn, and PyTorch.
- Bioinformatics libraries including Biopython for sequence analysis.
- Genome annotation and pathway analysis tools for biological feature extraction.
- Database resources including KEGG and UniProt for functional and metabolic information.
- Jupyter-based computational environments for workflow development and analysis.

4.2 Dataset Construction

The evaluation dataset is generated from experimentally measured microbial growth profiles combined with genome-derived biological features. Each sample represents a unique strain-condition combination:

$$Sample_i = (Strain_i, Condition_i, Genome_i, Growth_i) \tag{6}$$

The long-term objective of the framework is to scale this dataset toward one million experimentally characterized strain-condition combinations by evaluating approximately 1,000 microbial strains across 1,000 environmental conditions.

4.3 Performance Evaluation Metrics

The predictive performance of the developed models is evaluated using quantitative statistical metrics. For continuous growth prediction tasks, evaluation metrics include:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Coefficient of Determination (R^2)
- Pearson correlation coefficient between predicted and experimentally measured growth parameters

Model robustness is further evaluated through:

- Cross-validation across experimental conditions.
- Validation using independent wet-laboratory site measurements.
- Evaluation of prediction performance across different microbial strains and environmental conditions.

The combination of experimental reproducibility assessment and computational evaluation provides a comprehensive framework for determining whether genome-informed machine learning models can accurately predict microbial growth behavior.

5 Results

This section presents results to date from the experimental-computational framework under development. The experimental growth-characterization component is substantially complete for the pilot panel. The computational component, comprising genome annotation, pathway feature extraction, and predictive modeling, is at an active development and validation stage: a genome annotation pipeline has been installed and is undergoing preliminary testing, a modular alternative pipeline has been designed to address licensing and infrastructure constraints, and a systematic benchmark comparing both is planned. Predictive modeling of growth from genomic features has not yet been performed and is reported here as planned future work.

5.1 Microbial Growth Characterization Across the Pilot Panel

Growth was characterized for 25 bacterial reference strains spanning four phyla (Proteobacteria, Firmicutes, Actinobacteria, Deinococcus-Thermus) across 11 growth media, under aerobic and anaerobic atmospheres and two temperature regimes. Growth was measured by optical density (OD600) monitoring over time, and individual growth curves were fit to a sigmoid model to extract four kinetic parameters per well: maximum growth rate, maximum optical density (MaxOD, a proxy for biomass yield), lag time, and area under the curve (AUC). The resulting tidy dataset comprises 25 well-level observations across 6 columns.

Representative growth curves are shown in Figure 2.

5.2 Oxygen Availability Is the Dominant Determinant of Growth

Comparison of aerobic and anaerobic conditions across the panel showed markedly higher growth under aerobic conditions, with approximately 3-fold higher MaxOD and approximately 5-fold higher maximum growth rate relative to anaerobic conditions. This is consistent with the substantially greater ATP yield of aerobic respiration (approximately 32 ATP per glucose) relative to fermentative metabolism (approximately 2 to 4 ATP per glucose).

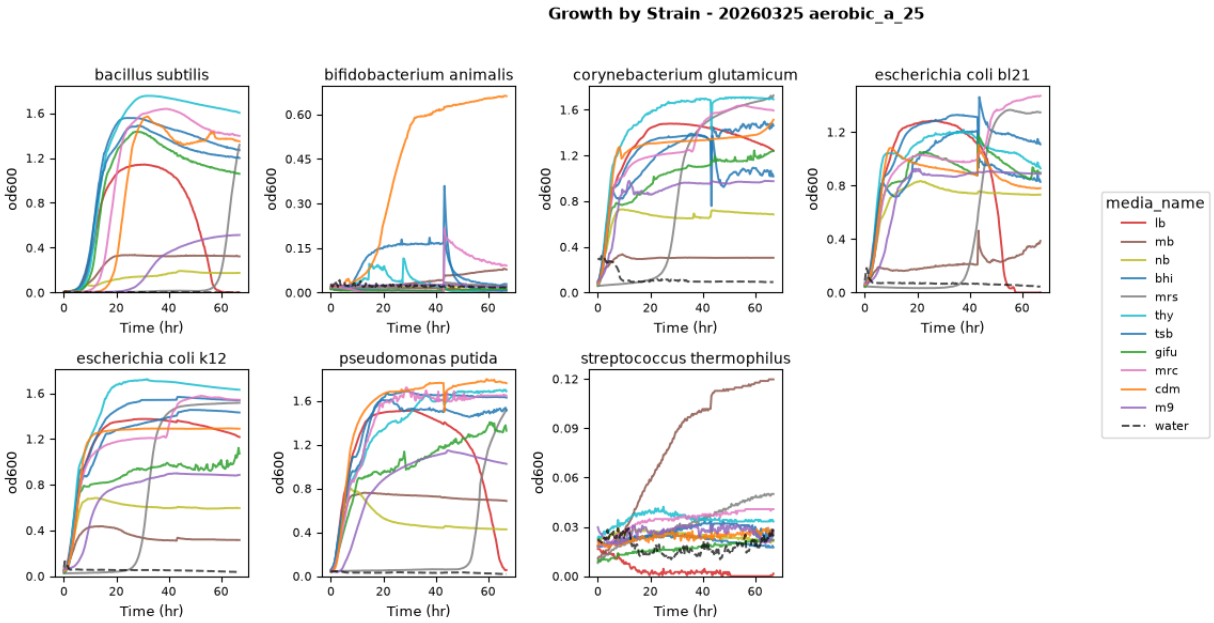


Figure 2 Microbial growth curves obtained from standardized cultivation experiments across the 25-strain pilot panel, showing kinetic diversity across strains, media, and atmospheric conditions.

Two strains, *Clostridium sporogenes* and *Lactiplantibacillus plantarum*, showed better growth under anaerobic than aerobic conditions, consistent with their known metabolic strategies and supporting that the observed oxygen effect reflects genuine strain-level physiology rather than a systematic assay artifact. Oxygen-requirement classification of the full panel identified 16 obligate aerobes, 5 facultative strains, 0 non-growing strains, and 3 obligate anaerobes.

5.3 Media Composition Shapes Biomass Yield and Growth Rate Differently

Growth medium was a strong secondary determinant of outcome. The richest media, MRC and TH+Y, produced the highest mean biomass yield (mean MaxOD of approximately 0.96 and 0.94, respectively), while Marine Broth and Nutrient Broth produced the lowest (mean MaxOD below 0.30). LB medium produced the highest mean growth rate (approximately 2.85 h^{-1}) among media tested, indicating that the medium that maximizes growth rate is not the medium that maximizes biomass yield.

5.4 Growth Rate and Biomass Yield Are Weakly Correlated and Should Be Treated as Distinct Prediction Targets

Across the panel, growth rate and MaxOD showed a weak linear relationship (Pearson $r \approx 0.24$) but a moderate monotonic relationship (Spearman $\rho \approx 0.65$), indicating a non-linear, rank-preserving trade-off between rate and yield consistent with proteome allocation constraints on aerobic metabolism. In contrast, MaxOD and AUC were strongly collinear ($r \approx 0.82$), reflecting their shared dependence on total biomass accumulated. This dissociation motivates treating growth rate and biomass yield as separate prediction targets in downstream modeling rather than as a single composite growth signal.

Strain-level variability in growth rate also differed substantially. *Bacillus subtilis* showed the greatest variability, including outlier values under aerobic conditions falling outside the plotted range in addition to visible anaerobic outliers, consistent with its capacity to use a broad range of carbon sources. *Pseudomonas putida* showed no outliers and the tightest growth-rate

distribution among strains examined, consistent with its characterization as a metabolically canalized laboratory strain.

5.5 Experimental Reproducibility Across Sites

The experimental workflow is designed to generate comparable growth profiles across independent wet-laboratory sites. Quantitative cross-site comparison of extracted growth parameters is ongoing and will be reported upon completion of the multi-site validation phase.

5.6 Data Quality Control

Systematic quality control identified two categories of measurement artifact requiring exclusion prior to downstream analysis: corrupted lag-time values (affecting approximately 10 wells, with erroneous magnitudes up to 8.7×10^{15} hours) and growth-rate values exceeding 13 h^{-1}, which were identified as sigmoid curve-fitting artifacts rather than biological signal. An automated filtering script was developed to quarantine flagged observations by exact match on experiment, date, strain, and medium, cleaning the data while preserving an auditable record of all exclusions.

5.7 Strain Metadata Standardization Pipeline

An automated pipeline was developed to acquire and standardize strain metadata from public repositories, including NCBI and ATCC. The pipeline resolved metadata for all 25 strains in the pilot panel with zero failures, and includes a runtime name-normalization layer that maps inconsistent raw strain identifiers to canonical names without modifying source data, supporting reproducible strain resolution as the panel scales toward the initiative's 1,000-strain target.

5.8 Genome Annotation Pipeline: Installation and Preliminary Testing

Genome-to-feature extraction requires converting each strain's genome assembly into a numeric representation of metabolic capability suitable for machine learning. The DRAM (Distilled and Refined Annotation of Metabolism) annotation pipeline was selected for its capacity to derive KEGG module completeness alongside CAZyme and peptidase annotations from a single tool.

DRAM v1 was successfully installed on the Digital Research Alliance of Canada's Nibi high-performance computing cluster. Installation required resolving several upstream issues independent of the present pipeline, including a documented syntax error in the DRAM-bio package's setup script and a stale external download URL affecting the dbCAN reference database. As the laboratory's KEGG license had not yet been obtained at the time of this work, the annotation database was configured to use KOfam, KEGG's freely available Hidden Markov Model profile database, in place of the license-gated UniRef90 database. Preliminary annotation testing on a small subset of the pilot genomes is in progress; full-panel annotation and parsing of pathway completeness output had not been completed at the time of this report.

5.9 A Modular, License-Independent Alternative Pipeline

In parallel, a modular alternative pipeline was designed to address two constraints encountered during DRAM deployment: the laboratory's lack of a KEGG license, and repeated infrastructure fragility during DRAM installation on shared HPC resources. The alternative pipeline decomposes annotation into three independent stages: gene calling with Pyrodigal, KEGG Ortholog assignment via a PyHMMER-based implementation of the KOfam search against the same KOfam Hidden Markov Model profiles used by DRAM's license-free configuration, and module completeness scoring via the EMBL-EBI `kegg-pathways-completeness-tool`, which computes completeness from KEGG module network structure rather than flat presence-absence counting. CheckM2

genome-quality scores are retained as a covariate to distinguish true gene absence from genome incompleteness. An optional parallel branch using eggNOG-mapper was also designed to recover CAZyme annotations, which map to carbohydrate-degradation capability and are mechanistically relevant given that the pilot panel spans 11 compositionally distinct media.

Because every component of this pipeline relies on publicly available reference data, it removes the KEGG licensing dependency entirely while preserving the same annotation models and module-completeness rigor as DRAM's license-free configuration. Its modular design also allows any single stage to be replaced independently as the project scales, consistent with the project's config-over-code architectural principle.

5.10 Planned Benchmark of Annotation Pipelines

To determine which annotation approach best serves the project at scale, a systematic benchmark has been designed comparing three configurations: the installed DRAM v1 pipeline under KOfam-only operation (Plan A), the modular license-independent pipeline described above (Plan B), and the current DRAM v2 public beta release (Plan C). The benchmark will evaluate the three approaches on a common subset of pilot genomes along four axes: pathway and annotation feature yield, pairwise concordance of module completeness scores between pipelines, computational cost (wall-clock time and peak memory), and setup and maintenance complexity. This benchmark had not yet been executed at the time of this report.

5.11 Future Evaluation and Benchmarking of Predictive Models

Following completion of genome annotation and feature extraction, strain-level pathway completeness features will be joined with experimentally measured growth parameters to train predictive models of microbial growth phenotype from genomic and environmental inputs. Planned modeling will benchmark regularized linear and logistic regression, gradient-boosted trees, and random forests, with validation performed using strain-level splits stratified by phylogenetic distance to assess generalization to unseen organisms. Given the pilot panel size (n = 25 strains), this evaluation will be reported as a feasibility assessment rather than a definitive performance benchmark, with model performance reported as a function of genomic distance to the nearest training strain. Planned evaluation metrics include mean absolute error, root mean square error, coefficient of determination (R^2), and correlation between predicted and observed growth parameters, benchmarked against baseline models using environmental variables alone.

6 Discussion

The preliminary results demonstrate the successful generation of standardized microbial growth profiles and the establishment of the initial computational infrastructure required for large-scale genotype-to-phenotype modeling. Although predictive modeling and metabolic pathway integration are not yet complete, the current findings provide important validation of the experimental and computational components that form the foundation of the proposed framework.

The observed microbial growth curves indicate that the experimental workflow is capable of capturing strain-specific growth behaviors under controlled cultivation conditions. Differences in growth dynamics between strains are expected, as microbial growth is influenced by variations in metabolic capacity, nutrient utilization pathways, regulatory mechanisms, and environmental adaptation. The ability to generate distinguishable growth profiles is an essential prerequisite for developing predictive models, as machine learning approaches require sufficiently diverse phenotypic variation to identify relationships between genomic features, environmental conditions, and microbial performance.

A major consideration in this work is experimental reproducibility. Since microbial growth measurements can be affected by laboratory-specific factors, including inoculation procedures, measurement equipment, and environmental fluctuations, cross-site validation is necessary to determine whether observed differences represent true biological variation rather than experimental noise. The ongoing comparison between independent wet-laboratory sites will provide an assessment of dataset reliability and determine whether the generated measurements are suitable for machine learning applications.

The computational progress achieved during this phase addresses a major limitation in previous microbial prediction studies: the lack of standardized integration between experimental phenotype data and biological features. The completed metadata retrieval and cleaning pipeline establishes a consistent data foundation, while the ongoing implementation of DRAM enables conversion of genome sequences into interpretable metabolic pathway representations. The resulting strain-by-pathway matrix is expected to provide biologically meaningful features that can complement environmental variables and experimentally measured growth parameters.

At the current stage, direct comparison with predictive modeling benchmarks is not yet possible because model training and evaluation have not been completed. By the end of this program, the work will determine whether incorporating metabolic pathway features improves growth prediction compared with baseline models using environmental information alone. Additionally, the final framework will be evaluated based on its ability to generalize across microbial strains, cultivation conditions, and independent experimental measurements.

No unexpected biological findings have been identified during this preliminary phase. However, the generation of large-scale, reproducible microbial growth datasets may reveal previously uncharacterized relationships between metabolic capabilities and environmental adaptation. These discoveries will depend on the successful integration of experimental phenotypes with genomic and metabolic features during the subsequent modeling phase.

7 Conclusion

This work presents the development of an integrated experimental-computational framework for predictive microbial growth modeling by combining standardized wet-laboratory characterization, genomic data processing, metabolic pathway reconstruction, and machine learning-based prediction. In the current phase, the experimental workflow has successfully generated preliminary microbial growth profiles, while an automated computational pipeline has been established for strain metadata retrieval, data cleaning, and genome information processing from resources including NCBI and ATCC. The ongoing implementation of DRAM enables the extraction of metabolic pathway features from genome sequences, providing the foundation for constructing strain-by-pathway representations that will be integrated with experimentally measured growth phenotypes. Although predictive model development and performance evaluation are ongoing, the current progress demonstrates the feasibility of establishing a scalable and reproducible platform for connecting microbial genotype, metabolism, and phenotype. The long-term goal of this framework is to enable large-scale microbial growth prediction across diverse strains and environmental conditions, reducing experimental screening requirements and accelerating applications in metabolic engineering, biotechnology, and predictive microbiology.