Tools for interfacing with existing databases (LivMat D4.1)

Luca Battiston

September 1, 2025

Abstract

Deliverable 4.1 outlines the development of a computational pipeline for generating numerical representations of bacterial strains to enable machine learning (ML) applications in synthetic microbial consortia design. We interface with established biological databases—namely BioCyc and BacDive—to extract genomic, metabolic, and physiological data. We present three distinct digital representation frameworks: a Genetic Representation (GR) based on gene presence-absence, a Metabolic Representation (MR) derived from metabolic pathways, and a Physiological Representation (PhyR) encoding interpretable macroscopic traits. These vectorized representations transform categorical biological data into an ML-ready format, providing a foundational toolset for subsequent predictive modeling within the LivMat project. All code and examples are made publicly available.

1 Introduction

During the initial phase of this project, we conducted a survey of publicly available biological databases to identify robust sources of strain-specific data. The primary objective was to establish a method for creating a reliable digital representations of microbial species. This report details the results of this exploration and presents a structured computational framework for converting categorical biological data from these databases into numerical feature vectors, suitable for informing ML models tasked with designing and optimizing microbial consortia.

2 Explored Databases

This section highlights the key databases investigated for their potential to provide the requisite strain-specific features for digital representation.

2.1 BioCyc Database

The BioCyc database collection is a comprehensive resource that integrates genomic, metabolic and regulatory informations of thousands of organisms. It provides annotations of genes, enzymes, reactions and metabolic pathways, making it particularly valuable for systems biology and, in our case, provides a reliable source of data for individual microbial species.

When looking for a specific strain, the BioCyc database might encounter several matching results. The website includes a Tier system and also a reference genome flag in order to indicate the quality of annotated data for the found strains. Each strain may be included in Tier 1, 2 or 3 which means:

- **Tier 1**: Has extensive and high-quality information and has been manually curated by experts using the Pathway Tools software [1, 2].
- **Tier 2**: May contain a mix of computationally predicted pathways plus some expert-reviewed with moderate manual curation.
- **Tier 3**: Computationally generated databases created automatically by running PathoLogic (a Pathway Tools component) on the genome secuence. No manual curation

The **reference genome flag** indicates that the database has been designated (often by NCBI/RefSeq or the research community) as the standard representative genome of that species. When a strain has this flag, it usually denotes high quality annotations that serve as main comparison point when studying different strains of the same species.

2.2 The BacDive database

The Bacterial Diversity Metadatabase (BacDive) is a comprehensive, publicly accessible resource providing structured phenotypic and taxonomic information for a vast array of bacterial strains. BacDive serves as a central hub for curated experimental data sourced from scientific literature, culture collection catalogs, and direct submissions from researchers.

Its value lies in its rigorous standardization of complex phenotypic data into searchable and comparable fields. The database provides detailed information on a wide range of traits, including but not limited to:

- Morphology and physiology (e.g., cell shape, dimensions, Gram staining).
- Culture and growth conditions (e.g., optimal temperature, pH range, salinity tolerance).
- Metabolism (e.g., carbon source utilization, enzyme production, oxygen requirement).
- Environmental isolation sources and biogeography.

2.3 Protist Interaction Database (PIDA)

The Protist Interaction Database is available online in [3] and can be easely downloaded as a `.xlsx` or `.tsv` file. The data set consists of a clasification of two organisms down to the species level (if possible) and their respective ecological interaction. These interactins are organized in the following categories:

- **Parasitism**

- **Predation**
- **Symbiosis** (either Mutualism or Commensalism)
- **Unresolved**

The **symbiosis** category in some cases also has further information that describes of symbiotic relationship in detail (See Figure 1).

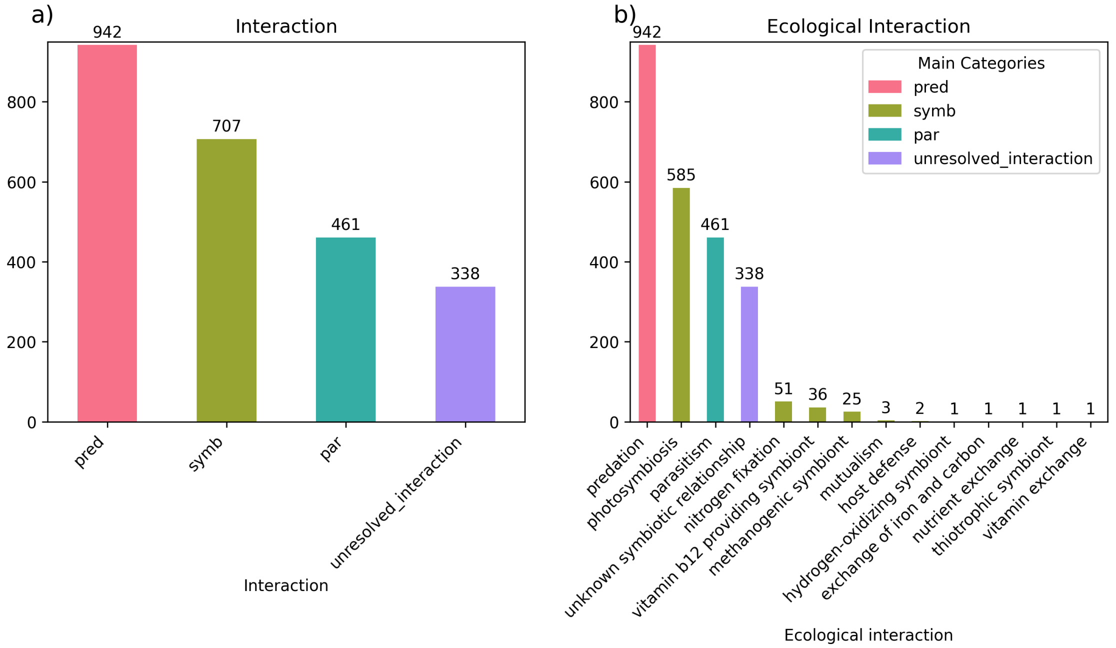

Figure 1: Histograms of the PIDA database showing the ammount of each interactions that are registered in the dataset. Figure **a)** show the interactions in the basic categories and Figure **b)** expands the interactions with further specification of the symbiotic interactions.

The `.tsv` file consist in a table of 2447 rows and 22 columns, where rows correspond to pairwaise organism intecatcions and columns indicate interactions and organizms taxonomy. In Figure 1 we can see most of the interactions involve predation. Less than half of the rows (exactly 706 rows) correspond to symbiotic interactions. Interestingly, 585 of these rows are classified as a photosymbiosis type of interaction (an interaction of symbiotic type where one of the organisms is capable of performing photosynthesis).

As shown in Figure 1, the data set includes unresolved pairwise interactions but also includes interactions where the complete one or both of the species have an incomplete taxonomical classifications. If these cases are removed from the dataset in order to take into account only complete classifications and resolved interactions, the dataset reduces to 581 predation interactions, 510 parasitic and 126 symbiotic.

3 Digital bacterial representation

When developing Machine Learning (ML) models, its is essential to understand the inner workings of the most basic algorithms. By doing so, we gain an understanding of how these algorithms work, how to debug them, how to tune the hyperparameters and also the importance of having proper high-quality data that is suitable for wichever model we choose to train.

In the databases explored in the previous sections (particularly in BioCyc), there is a high availability of categorical data for bacterial species. Most ML models require a numerical input from which the model generates a prediction, calculates a loss values and updates the model parameters. These tasks are clearly not achievable when working with categorical data, which poses the need of developing a preprocessing step on the available data. With this idea in mind we aimed to create a digital representation of selected bacterial traits by producing a vector-embeding of the categorical data in order to be able to feed this data to the ML models to be developed so that the models are informed about the specific bacterial strains that integrate the consortia of interest. We developed 3 different representations that will be described below.

3.1 Genetic and Methabolic Representations (GR & MR)

Microbial cells, though unicellular, are highly complex living systems that can be understood through complementary analytical frameworks. A genetic framework defines an organism by its genome, the complete set of genes encoded within its DNA, which provides the blueprint for its capabilities. Conversely, a metabolic framework views the cell as a integrated biochemical reactor, where coordinated networks of enzymes and metabolic pathways transform extracellular nutrients into energy, biomass, and valuable metabolic products.

As previously mentioned, the BioCyc database provides structured, categorical biological data for a wide range of bacterial strains. This data includes, for each strain, a comprehensive list of annotated genes and their associated metabolic pathways. The current section details the construction of a Genetic Representation (GR) vector for each bacterial strain. It is important to note that an analogous procedure can be applied to the metabolic pathway data to construct a Metabolic Representation (MR), leveraging the same underlying computational framework.

For each individual strain, the process begins by extracting a gene list from BioCyc in the form of `.tsv` file. This file contains several columns, including a description of the gene's functional product and a unique identifier for each gene known as a locus tag (e.g., `RSXXXXX`). For the purpose of constructing the GR, we are only interested in the presence or absence of specific genes. Therefore, we extract only the complete set of locus tags for a given strain, which serves as the definitive record of its genetic repertoire.

Formally, consider a set of n_{sp} bacterial species, denoted by $\{s_i\}_{i=1}^{n_{\text{sp}}}$. The total number of unique genes across all species in this set is n_g. We represent the k-th bacterial strain by a binary vector $\vec{s}_k^{\,\text{GR}} \in \{0,1\}^{n_g}$, where each component (or feature) in the vector corresponds to a unique gene from the universal set. The value of a component is 1 if the strain possesses the corresponding gene, and 0 if it does not. This method, known as **one-hot encoding** or creating a presence-absence matrix, is a foundational technique in machine learning for converting categorical data into a numerical format, treating each category as an independent binary feature.

For this propose, we have developed a custom Python script that acts as an interface between the raw BioCyc data and a downstream machine learning pipeline. The algorithm takes in an arbitrary number of BioCyc `.tsv` files, compiles a global set of all unique locus tags, and subsequently generates the corresponding $\vec{s}_k^{\mathrm{GR}}$ vector for each input strain. The output is a feature matrix $\mathbf{S}^{\mathrm{GR}} = [\vec{s}_1^{\mathrm{GR}}, \vec{s}_2^{\mathrm{GR}}, \ldots, \vec{s}_{n_{\mathrm{sp}}}^{\mathrm{GR}}]^T$, which is ready to be used by standard machine learning algorithms. We term these vectors the Genetic Representation (GR) of the bacterial species.

A schematic overview of this process is provided in Figure 2.

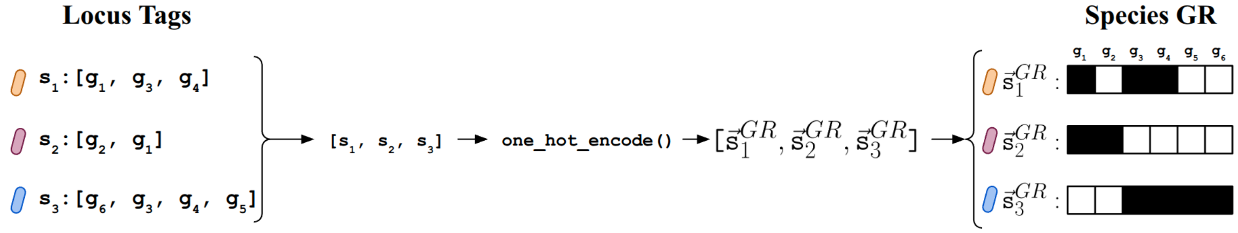

Figure 2: Schematic of the Genetic Representation (GR) vector construction process. The algorithm processes the locus tags from three input species, generates a unified set of all unique genes, and outputs a binary matrix. A filled box (1) indicates the presence of a gene in a species' genome; an empty box (0) indicates its absence.

As previously mentioned, the computational framework developed for the Genetic Representation (GR) can be directly applied to construct the Metabolic Representation (MR). This creates a significant strategic opportunity to integrate our ML efforts with the metabolic modeling activities conducted in Work Package 1 (WP1).

Genome-scale metabolic models (GEMs) are computational representations of the entire metabolic network of an organism, comprising the full set of biochemical reactions it can catalyze. Our pipeline can be seamlessly adapted to utilize these reaction sets as the categorical input for vectorization. Instead of encoding the presence of genes, the MR would encode the presence of metabolic reactions, thereby creating a numerical profile of the strain's metabolic capabilities.

We belive this synergy is critical for building predictive models that are both biologically interpretable and directly relevant to the project's aims.

3.2 Physiological Representation (PhyR)

While the Genetic (GR) and Metabolic (MR) Representations provide high-dimensional digital profiles of bacterial strains, they can lack immediate physiological interpretability. To bridge the gap between genomic data and macroscopic, observable traits, we developed a third representation: the Physiological Representation (PhyR). The core objective of the PhyR is to encode key physiological characteristics that are directly measurable, biologically meaningful, and thus more readily interpretable.

To construct the PhyR, we extended our data sourcing beyond genomic databases to include scientific literature and curated biological data from BacDive that contain experimental measurements of macroscopic features. We began by collecting a foundational set of general physiological traits, including morphology (shape, dimensions), ideal growth conditions (temperature, pH), motility, oxygen tolerance, pigment production, and Gram stain classification. Table 1 presents this manually curated physiological feature set for the two foundational bacterial strains in this project: *Synechocystis sp.* PCC 6803 and *Pseudomonas taiwanensis* VLB120.

Feature	***Synechocystis***	***P. taiwanensis***
Shape	Spherical[4]	Rod-shaped
Length (μm)	0.7 – 8[4]	1.0 – 2.2
Width (μm)	0.7 – 8[4]	0.7 – 1.0
Motile	Yes[5]	Yes
Pigment	Yes[6]	No
Ideal Temp. (°C)	32 – 38[7]	30 – 37
Ideal pH	7.0 – 8.5[8]	6.0 – 8.0
O_2 tolerance	Aerobe[9]	Aerobe
Gram stain	Negative	Negative

Table 1: Curated physiological traits for the primary bacterial strains under investigation. Data cited from literature; uncited entries were sourced from the BacDive database.

The data in Table 1 consists of both continuous numerical ranges (e.g., temperature, pH) and categorical values (e.g., shape, motility). To render this data suitable for machine learning, a multi-step encoding procedure was applied:

- **Categorical Features:** These were converted using a one-hot encoding scheme, similar to the method used for the GR. For example, the 'Shape' category, which can be Spherical, Rod-shaped, or Spiral, was split into three separate binary features. A strain is assigned a value of 1 for the feature corresponding to its shape and 0 for all others.
- **Numerical Features:** For features reported as a range (e.g., Length: 0.7–8 μm), we calculated the midpoint to obtain a single, representative scalar value (e.g., (0.7 + 8)/2 = 4.35 μm). This provides a consistent numerical summary of the range.
- **Binary Features:** Traits with simple yes/no or present/absent responses (e.g., Motile, Pigment) were directly encoded as 1 or 0.

The result of this encoding process is the machine-readable feature matrix presented in Table 2. Each bacterial strain is now represented by a fixed-length numerical vector, $\vec{s}_k^{\,\mathrm{PhyR}}$, where each dimension corresponds to a specific, interpretable physiological trait.

This transformation yields a lower-dimensional yet highly interpretable digital representation of our bacterial strains. The PhyR vectors provides features that are not only

Feature	***Synechocystis***	***P. taiwanensis***
Shape: Spherical	1	0
Shape: Rod-shaped	0	1
Shape: Spiral	0	0
Length (μm)	4.35	1.60
Width (μm)	4.35	0.85
Motile	1	1
Pigment	1	0
Ideal Temp. (°C)	35.0	33.5
Ideal pH	7.75	7.00
O_2: Aerobe	1	1
O_2: Anaerobe	0	0
Gram stain: Positive	0	0
Gram stain: Negative	1	1

Table 2: Machine-learning-ready Physiological Representation (PhyR) vectors for the bacterial strains. Categorical traits have been one-hot encoded, and numerical ranges have been summarized by their midpoint.

consumable by ML algorithms but also directly map to tangible, experimental physiological properties.

4 Conclusion and Future Work

This deliverable has established a computational pipeline for interfacing with biological databases and generating machine-learning-ready representations of microbial species. We have created a structured methodology for encoding biological raw data from databases into a numerical language that algorithms can process. The three representations—Genetic (GR), Metabolic (MR), and Physiological (PhyR), provide complementary views of a bacterial strain, from its genomic blueprint and metabolic potential to its macroscopic, observable traits.

The natural progression of this work is the development and training of ML models that consume these representations to predict community0level functions. The integration of the MR with genome-scale metabolic models (GEMs) from WP1 represents a particularly powerful synergy, potentially enabling models that predict functional outcomes from metabolic network structure. Furthermore, the pipeline is designed to be extensible; as new data sources from other Work Packages or relevant traits are identified, they can be incorporated into this flexible framework.

Data and code availability

The Python code for generating Genetic and Metabolic Representations, along with documented examples of its application is availabe at:

https://github.com/LucaBattt/LivMat-WP4-Deliverables/tree/main

References

[1] Romero P. Karp P.D. Paley S. “The Pathway Tools software”. In: *Bioinformatics.* (2002). DOI: doi:10.1093/bioinformatics/18.suppl_1.s225.

[2] Peter D et al. Karp. “Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.” In: *Briefings in bioinformatics vol. 11,1* (2010). DOI: doi:10.1093/bib/bbp043.

[3] GitHub repository with the Protist Interaction Database.

[4] Malihe Mehdizadeh Allaf and Hassan Peerhossaini. “Cyanobacteria: Model Microorganisms and Beyond”. In: *Microorganisms* 10.4 (2022). ISSN: 2076-2607. DOI: 10.3390/microorganisms10040696. URL: https://www.mdpi.com/2076-2607/10/4/696.

[5] Devaki Bhaya, Akiko Takahashi, and Arthur R. Grossman. “Light regulation of type IV pilus-dependent motility by chemosensor-like elements in ¡i¿Synechocystis¡/i¿ PCC6803”. In: *Proceedings of the National Academy of Sciences* 98.13 (2001). DOI: 10.1073/pnas.131201098.

[6] Nicole Tandeau de Marsac. “Phycobiliproteins and phycobilisomes: the early observations.” In: *Photosynthesis Research 76, 193–205 (2003)* (). DOI: https://doi.org/10.1023/A:1024954911473.

[7] Jan Červený et al. “Mechanisms of High Temperature Resistance of Synechocystis sp. PCC 6803: An Impact of Histidine Kinase 34”. In: *Life* 5.1 (2015), pp. 676–699. ISSN: 2075-1729. DOI: 10.3390/life5010676. URL: https://www.mdpi.com/2075-1729/5/1/676.

[8] Thorsten Heidorn et al. “Chapter Twenty-Four - Synthetic Biology in Cyanobacteria: Engineering and Analyzing Novel Functions”. In: *Synthetic Biology, Part A*. Ed. by Chris Voigt. Vol. 497. Methods in Enzymology. Academic Press, 2011, pp. 539–579. DOI: https://doi.org/10.1016/B978-0-12-385075-1.00024-X. URL: https://www.sciencedirect.com/science/article/pii/B978012385075100024X.

[9] Fernando Baquero and Marc Lemonnier. “Generational coexistence and ancestor’s inhibition in bacterial populations”. In: *FEMS Microbiology Reviews* 33.5 (2009). DOI: 10.1111/j.1574-6976.2009.00184.x.