VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria

Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS

TRABAJO DE FIN DE MÁSTER

EXPLORATION OF METABOLIC FUNCTIONS IN RÍO TINTO MARS-ANALOG

AUTOR/A:
GUILLERMO CLIMENT GARGALLO

TUTOR/A/S/AS:
PABLO JORGE CARBONELL CORTÉS
FRANCISCO MARTÍNEZ GIL
VALERO LAPARRA PÉREZ-MUELAS

SEPTIEMBRE, 2023

VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria

VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria ETSE-UV

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS

TRABAJO DE FIN DE MÁSTER

EXPLORATION OF METABOLIC FUNCTIONS IN RÍO TINTO MARS-ANALOG

AUTOR/A:
GUILLERMO CLIMENT GARGALLO

TUTOR/A/S/AS:
PABLO JORGE CARBONELL CORTÉS
FRANCISCO MARTÍNEZ GIL
VALERO LAPARRA PÉREZ-MUELAS

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

Abstract

NASA's Artemis program has revitalized our interest in the Red Planet. The ambitious project will enable the deployment of people on the Moon and its orbit once again, opening the door to a much imagined colony on Mars. Unsurprisingly, the program faces gargantuan challenges, especially those related to sustaining life for prolonged periods of time far from Earth. *In situ* resource utilization (ISRU) aims to solve these issues by using available resources in these settings.

One of the potential solutions for ISRU involves leveraging the natural capabilities of microorganisms for utilizing materials and producing goods, from food and pharmaceuticals to components and beyond. Due to the lack of samples from Mars, we need to rely on analog environments for extrapolation. Precisely, studying the microbial communities thriving in these analogs may shed some light not only on ways to sustain ISRU facilities, but on the origin of life.

In our case study, we focus on Río Tinto, a Mars analog in Spain with a rich subsurface community. By leveraging data from previous studies, we create a pipeline for exploring the metabolic potential of these communities from the perspective of a biofoundry on Mars. Our implemented pipeline also underpins how the optimization for the whole microbial community yields inconclusive results due to the limitations of the available data, although revealing viable production of several compounds of interest.

Keywords

Mars analog, Río Tinto, metabolic engineering, biofoundry, ISRU

Resumen

El programa Artemis de la NASA ha revitalizado nuestro interés en el planeta rojo. El ambicioso projecto permitirá mandar personas a la luna y su órbita una vez más, abriendo la puerta a una muy esperada colonia en Marte. No es de extrañar que este programa se encuentre con desafíos descomunales, especialmente los relacionados con mantener la vida por periodos prolongados lejos de la Tierra. La utilización de recursos *in situ* (ISRU por sus siglas en inglés) trata de solventar estos problemas utilizados los recursos disponibles en estos entornos.

Una de las posibles soluciones de ISRU supone aprovechar las capacidades naturales de los microorganismos para consumir materiales y producir productos de interés, desde comida y fármacos a componentes y demás tipo de útiles. Debido a la falta de muestras de Marte, necesitamos basarnos en entornos análogos para extrapolar las condiciones. Precisamente, estudiando las comunidades microbianas properando en estos análogos nos puede ayudar a arrojar algo de luz, no solo en maneras de mantener estos componentes de ISRU, sino también en el propio origen de la vida.

En nuestro caso, nos centramos en Río Tinto, un análogo de Marte en España con una rica comunidad subterránea. Aprovechando los datos de estudios previos, hemos podido crear una metodología para la exploración del potencial metabólico de estas comunidades desde la perspectiva de una biofactoría en Marte. No obstante, nuestra metodología también recalca que, debido a la limitación intrínseca que muestran los datos, los resultados obtenidos para la comunidad en su conjunto son inconclusivos. En cualquier caso, los resultados demuestran la viabilidad de la producción de algunos compuestos de interés.

Palabras clave

Análogos de Marte, Río Tinto, ingeniería metabólica, biofoundry, ISRU

Agradecimientos

Este trabajo ha sido un poco movido, lo tengo que confesar. Hace cuatro años que acabé todas las asignaturas del máster excepto el trabajo final de máster, y contacté con Paco y con Valero para hacerlo en Reinforcement Learning, eso de los robots que se mueven solos (o eso nos quieren hacer creer). Llegó una cosa que se llamó pandemia (no sé si os enterasteis), me desilusioné, cambié a otro tema que tenía que ver con proteínas y sus interacciones (cosas de esas biológicas) y me volví a desilusionar, porque yo soy así. No puedo evitarlo. Total, que acabé contactando con Pablo para que me ayudara a hacer un trabajo que tuviera parte de Ciencia de Datos y parte de Bioinformática, y si podía ser con un toque de Astrobiología, mejor. Es por esto que les agradezco su apoyo y guía durante este ajetreado viajecito.

No puedo olvidarme de agradecer a mis amigos por hacerme compañía en eso que por estas tierras se llama almuerzo pero se hace antes de la hora de comer. Tanto a los del pueblo como de a los de capi, gracias de verdad por esos cremaets en la terracita.

Obviamente, también tengo que agradecer a mis padres que me hayan aguantado cuando no me aguantaba ni yo, entre el calor y las prisas por acabar este trabajo. Como siempre, gracias por todo y por estar ahí. No habría llegado aquí sin vosotros. Y ya son tres trabajos finales!

Este es el apartado de agradecimientos de mi familia política, que los tengo que poner porque, si no, alguno se enfada y no me invita a fideuà! Gracias por esas charlas de desayuno que se alargan hasta la comida, tengo que confesar que me han alegrado más de una mañana.

También quería agradecer a esa personita que me entiende aunque no esté aquí y que sabe guiarme cuando estoy perdido y parece que no haya ningún rayo de esperanza. Gracias por los momentos tontos y por los regalos hechos a mano, me han sacado de algún que otro atolladero. Gracias, Mayra, por pasear conmigo sin rumbo fijo. Y como diría el famoso filósofo Sócatres: 'I'm Batman!'

Por último pero no por ello menos especial, quería agradecer a una bola de pelo que siempre me viene a saludar y a hacerme compañía (aunque yo también le hago compañía). Gracias de verdad, Moss, por estar ahí!

Space. It seems to go on forever. But then you get to the end and a gorilla starts throwin' barrels at you.

Philip J. Fry, Futurama, Space Pilot 3000

Contents

List of Figures

List of Tables

Acronyms

1 Introduction

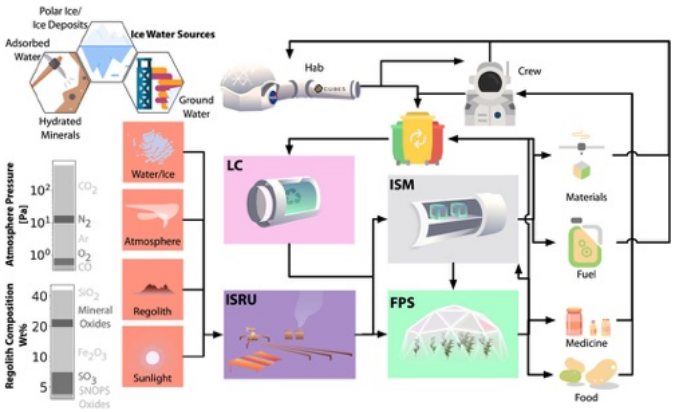
Mars has fascinated us since the dawn of our civilization. It is of no surprise, being our closest neighbor, potentially harboring life in the past (and present) and the focus of several missions, from the now extinct Union of Soviet Socialist Republics (USSR) to National Aeronautics and Space Administration (NASA). With the Artemis program, the next step to the Red Planet will become closer [2, 3, 4]. Indeed, NASA's plans for the near future involve a crewed mission to Mars in the 2030s [5]. These missions will set up the Gateway, providing a place to live and work, and supporting long-term science and human exploration on and around the Moon as well as, potentially, for future missions to Mars [2, 3]. The value of the Red Planet as a critical milestone in the space career is unquestionable, but how are we going to sustain human life on it?

Challenges. Missions beyond low Earth orbit (LEO) face gargantuan challenges. Currently, there is a lack of a coherent strategy and standards for the utilization and development of current technology towards a colony beyond LEO [6, 7]. However, there have been initiatives covering the broad goals of such endeavours [5, 4, 8]. Prolonged periods in deep space may imply less immediate access to life-sustaining elements readily available on Earth [4]. One major factor is the cost of supplies launched beyond LEO, which can range from 1,500 United States dollars (USD) (Falcon Heavy) to more than 30,000 USD (Minotaur IV) per kilogram [9, 10]. Additionally, Mars poses a highly challenging environment due to its extreme conditions compared to the ones humans have adapted to. For example, the enormous amount of radiation its surface receives [5] together with the presence of perchlorates in the regolith [11] imply a highly toxic environment for most of the organisms that are known. From a logistics perspective, water is a limiting resource, at least in the surface and away from the poles [5]. The lack of oxygen will also increase the complexity of return missions, since it is needed for the combustion of propellants [5, 8]. Last but not least, there is the ethical dimension to the establishment of a Mars colony: among other concerns, how can we be sure that there is no previous life on the planet and that our activities are not going to harm it? [12, 13]

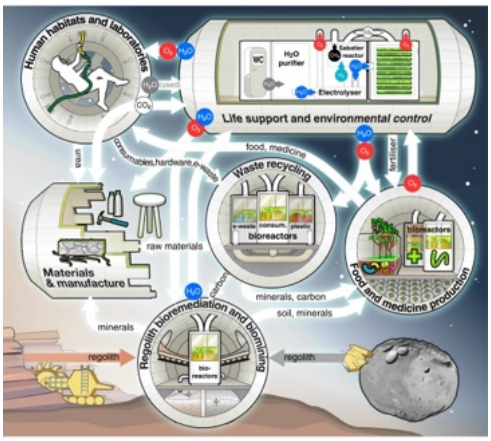
ISRU. The farther humans go into deep space, the more important it will be to generate products with local materials, a practice called *in situ* resource utilization (ISRU) or more generally as space resource utilisation (SRU) [4, 7, 8, 14]. ISRU is the practice of collection, processing,

storing and use of materials found or manufactured on other astronomical objects (the Moon, Mars, asteroids, etc.) that replace materials that would otherwise be brought from Earth [8]. During the last years, there has been an increasing interest on deploying ISRU facilities on the Moon and Mars (see Figure 1), specially given the importance of Artemis missions in the current space theatre [5, 2, 3, 4, 7, 11, 14, 15]. Making use of available resources on the surface and atmosphere of the bodies targeted by these missions will allow a longer autonomy and reduce launch costs [5, 4]. The potential to produce propellant, habitation and materials critical to support human life (e.g. water, oxygen, pharmaceuticals, food) positions this methodology at the forefront of space exploration (see Figure 1) [7, 11, 14, 16]. Proof of that is the testing of the Mars OXygen ISRU Experiment (MOXIE) payload on board NASA's Perseverance Rover produced oxygen from Mars' atmosphere by solid oxide electrolysis [17]. Other initiatives focus on the bioremediation of perchlorates (Martian Regolith Microbiome Inoculation Experiment (MaRMIE)) [18], the testing of pilot plants for self-sustained facilities in space (Micro-Ecological Life Support System Alternative (MELiSSA)) [19] or the growth of plants (Vegetable Production System (Veggie)) [20]. Software-centered approaches can also be found, such as the case of Deep Space Biology's Yotta, an explanaible Artificial Intelligence (AI) platform leveraging space biology data [21], or Lunco's Open Source software for Lunar colony engineering [22]. Confronting the promises of ISRU, there exist critical knowledge gaps that may hamper the development of such technologies on site [6, 7, 14, 16].

Biomanufactories. One of the potential solutions to ISRU on Mars are biomanufactories [11]. These facilities could supply the needs of a future colony by leveraging Mars natural resources by means of metabolic engineering, which involves the engineering and optimization of processes from single-cell to fermentation in order to increase the production rate of valuable chemicals for health, food, energy, materials and others [11, 16, 23]. For example, [11] propose a multi-module design in which the food, pharmaceuticals and components production as well as waste disposal and recycling are coupled to ISRU. By leveraging the metabolic capabilities of several species of microorganisms, this design tries to maximize resource utilization while minimizing launched cargo. The proposed biomanufactory consists on the following modules: ISRU, *in situ* manufacturing (ISM), food and pharmaceutical synthesis (FPS) and loop closure (LC). The ISRU component enables the production of food, pharmaceuticals and



(a) ISRU system proposed by [11] (CC BY 4.0)[1].



(b) ISRU system proposed by [16] (CC BY 4.0).

Figure 1: Two different schemes for ISRU on Mars (a, b) and the Moon (b). Both systems rely on living beings (microorganisms and/or plants) for extraction of essential components from the environment as well as production of goods. Special emphasis is applied to loop closure to reduce unnecessary waste. Among others, molecular oxygen and water will be critical to maintain habitability in these systems.

manufacturing of components by producing their input sources, mainly via biological fixation of atmospheric carbon and nitrogen, while regolith detoxification may provide a significant inventory for trace elements (Fe, K, P, S, etc.) [11, 14, 16, 18].

Biofoundries. An improved version of biomanufactories are biofoundries, which have the potential to enable a more automated ISRU by taking advantage of organisms adapted to Mars' conditions and resources. Biofoundries provide automation and analytics infrastructures to support the engineering of biological systems, enabling high-throughput scale for testing and, thus, increasing the solution space that can be explored [24]. Therefore, they intrinsically carry two main advantages over the rest of approaches: first, the reduction of launch costs thanks to the replicative capacity of biology [11], and second, their ability to adapt to different needs and conditions thanks to genetic engineering [24]. Additionally, well-built biofoundries will function mostly automatically, thanks to the Design-Build-Test-Learn (DBTL) cycle, allowing the crew to allocate more time on other important tasks (see Figure 2) [24, 25]. In the design step, the genetic constructs containing the genes of interest are created. These artifacts will later on be inoculated into the target organism(s) during the build stage. Afterwards, screening is performed during the test phase to select those with the desired trait (usually called phenotype). Finally, the

[1] https://creativecommons.org/licenses/by/4.0/

learn stage involves the analysis of the results and further decision making for the next cycles. There are many efforts involving the application of modeling and machine learning techniques in this stage [23, 24, 26, 27].

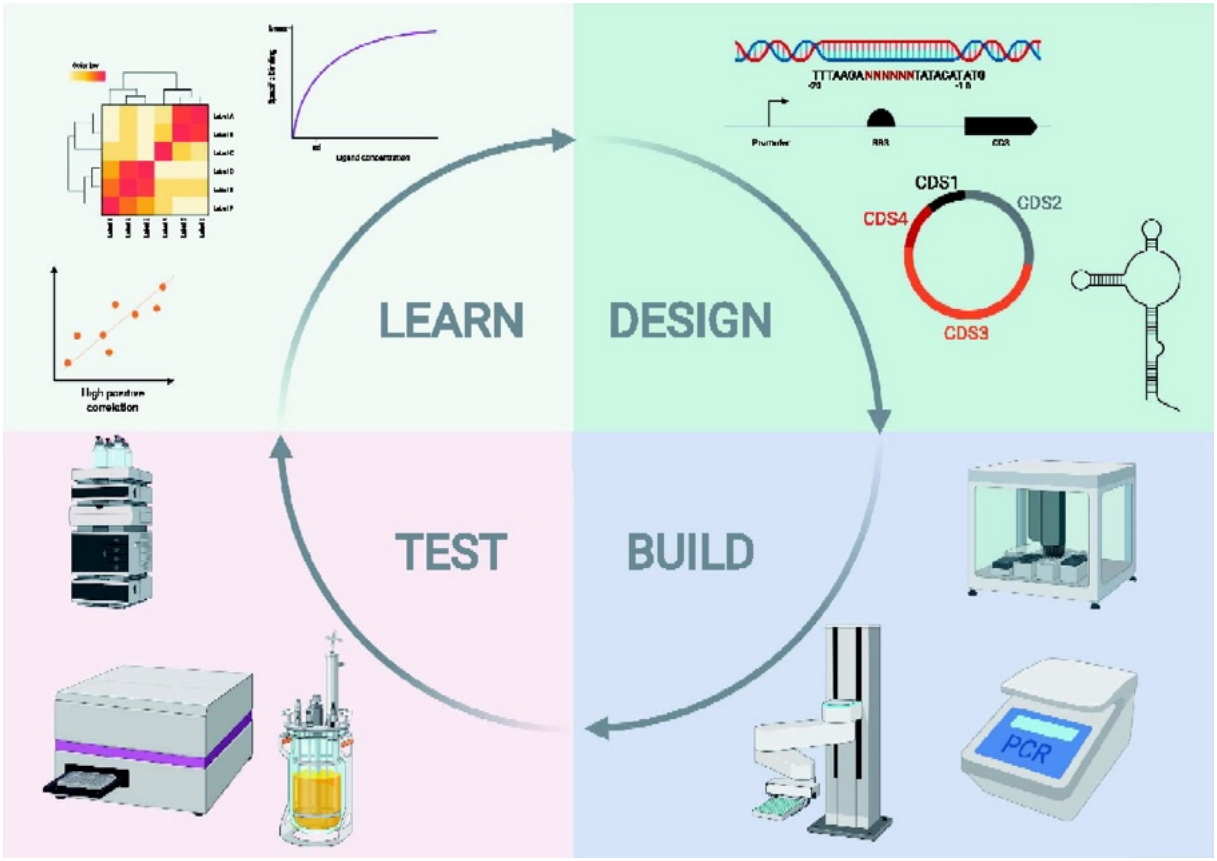


Figure 2: DBTL cycle adopted by biofoundries. By adapting to new conditions and learning from past data, biofoundries could overcome the disadvantages of biomanufactories in space, thus enabling a faster response to changing environments and needs. Figure taken from [24] (Open Government Licence v2.0)[2].

Mars analogs. Analog missions are field tests in locations that have physical similarities to the extreme space environments [28]. By allowing experimentation *in situ*, they help reduce costs and test a broad range of solutions that could be prohibitively expensive otherwise. Thanks to its properties, Río Tinto constitutes an analog environment of Mars on Earth [29] and can be studied to deepen our understanding of this mysterious planet and its conditions. Located in Huelva (Southwestern Spain), Río Tinto is an unusual extreme environment due to its acidity (mean pH 2.3), size (92 km long) and high concentration of heavy metals (Fe, Cu, Zn, As, etc.), as well as its high level of microbial diversity [29, 1, 30]. The oxygen content varies from saturation to complete anoxic conditions [29] and precipitates from Río Tinto's waters contain several iron sulfates and

[2] https://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

oxides, such as jarosite, which resemble those found at the landing site of the Opportunity Rover on Mars [29, 31]. These conditions contribute to the red coloration of its waters (see Figure 3).

(a) River view of Río Tinto. Courtesy of the author (CC BY 4.0).

(b) Detail of rocks in Río Tinto. Courtesy of the author (CC BY 4.0).

Figure 3: Río Tinto as a Mars analog. Thanks to its high acidity, metals concentration, similar mineralogy and thriving underground microbial communities, Río Tinto is studied to further understand the extremes of life on Earth and potentially on Mars.

Río Tinto can also be of industrial interest given the intrinsic Biotechnological potential of the enzymes catalyzing critical steps in this environment (see Figure 4 and 5) [32]. Although, its waters and deep subsurface have been thoroughly studied [1, 30], we decided to focus on the information related to underground communities for the availability of annotated genomes for several isolated species (see Table 1). A non-comprehensive list of metabolic functions studied in this community is depicted in Figure 5, where a bias can be observed for higher depth.

Biochemistry background. All living beings carry their genetic information in their DNA, a relatively stable molecule that can be passed down to their offspring. For the DNA to exert its function, it must be transcribed into RNA and then into proteins. From a functional perspective, both kinds of molecules have their own set of functions inside the cell, but we are going to focus on the proteins, specifically on a subset called enzymes. These enzymes are the catalysts of life: they accelerate chemical reactions necessary for living beings that would be otherwise extremely slow. This set of reactions is called metabolism and has three main objectives: the conversion of food into energy to run cellular processes; the conversion of food into building blocks for cellular components; and the elimination of metabolic wastes. In order to facilitate the exploration of the metabolism

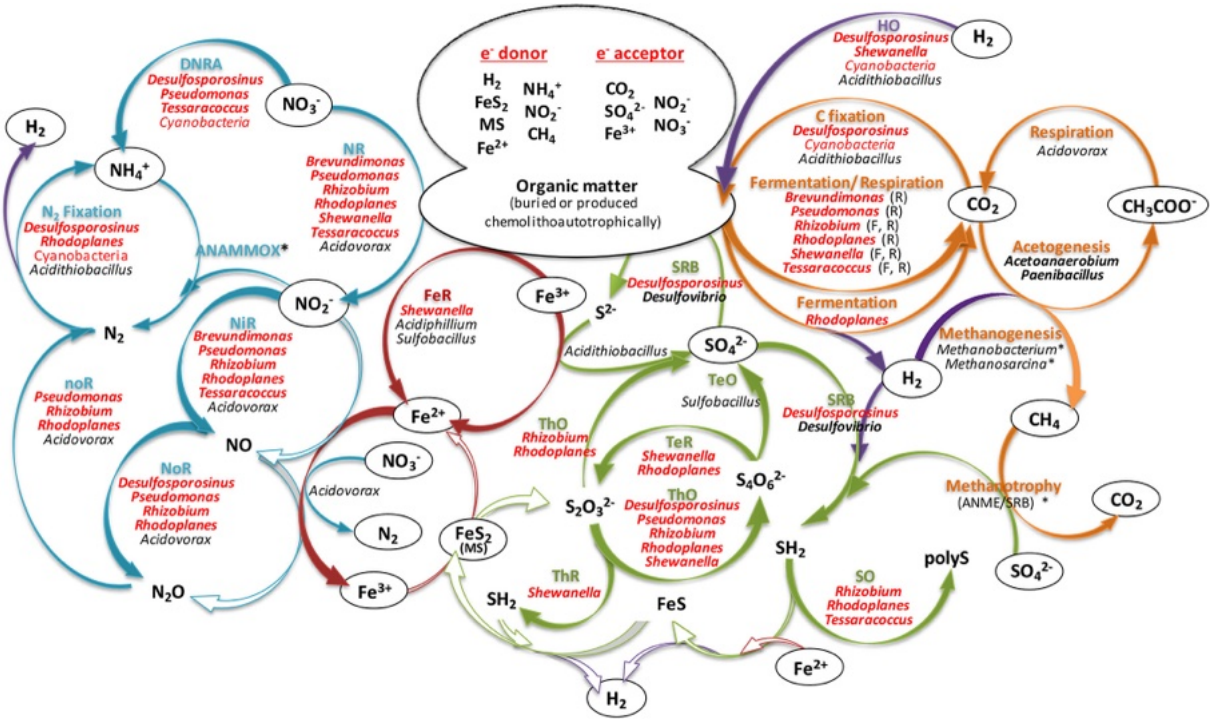


Figure 4: Geomicrobiological model of the C, H, N, S and Fe biogeochemical cycles operating in the deep subsurface of Río Tinto. In red, all isolated genera for which annotation is available and gene corresponding for the reaction has been identified. Some steps in the S (oxidation to sulfate) and C cycles (methane-related pathways and acetogenesis) are missing in the data, thus limiting the scope of the experiments. Figure taken from [1] (CC BY 4.0).

of living beings, a genome-scale metabolic model (GEM) can be created for a single species, thus allowing its *in silico* analysis and modification, potentially spotting any missing reactions and, thus, enzymes [33, 34]. Some methods even extend this methodology for working with groups of species (also called communities), such as the soil or gut microbiome [35, 36, 37].

1.1 Pipeline.

The methodology followed in this work combines both manual and programmatic steps. Figure 6 depicts the process from the selection of genomes to the launch of the RetroPath [38] and MICOM [35] tools for assessing the metabolic feasibility and predicting community growth, respectively. The pipeline starts with the preliminary literature review and selection of the environment and microbial species. Afterwards, annotated genomes are retrieved for the selected species from GenBank [39] and their corresponding metabolic models are created using ModelSEEDpy[3] [40]. This tool queries the RAST server [41] for the selected genomes in order to

[3] https://github.com/ModelSEED/ModelSEEDpy

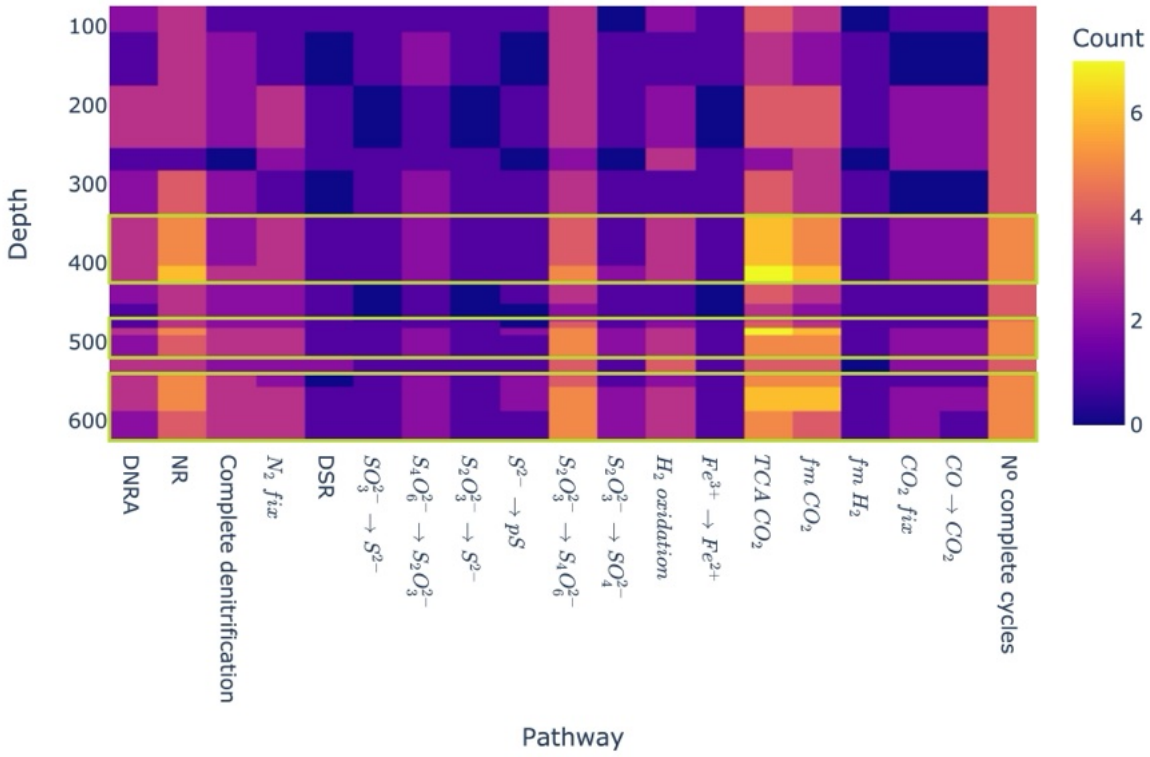


Figure 5: Metabolic functions extracted from supplementary information in [1]. Values represent the number of microbial species detected at different depths which have the potential of carry out key metabolic pathways of the C, H, N, S and Fe cycles. The depths at which all cycles are represented are highlighted in light green rectangles (see column N^o *of complete cycles*). A bias for complete cycles can be observed at higher depth. DNRA: dissimilatory nitrate reduction to ammonium; NR: nitrate reduction; N_2 fix: nitrogen fixation; DSR: dissimilatory sulfate reduction; $SO_3^{2-} \rightarrow S^{2-}$: sulfite reduction to sulfide; $S_4O_6^{2-} \rightarrow S_2O_3^{2-}$: tetrathionate reduction to thiosulfate; $S_2O_3^{2-} \rightarrow S^{2-}$: thiosulfate reduction to sulfide; $S^{2-} \rightarrow pS$: sulfide oxidation to polysulfides; $S_2O_3^{2-} \rightarrow S_4O_6^{2-}$: thiosulfate oxidation to tetrathionate; $S_2O_3^{2-} \rightarrow SO_4^{2-}$: complete thiosulfate oxidation to sulfate; H_2 oxidation: hydrogen oxidation; $Fe^{3+} \rightarrow Fe^{2+}$: iron reduction; TCA CO_2: carbon dioxide production by respiration; fm CO_2: carbon dioxide production by fermentation; fm H_2: hydrogen production by fermentation; CO_2 fix: carbon fixation; $CO \rightarrow CO_2$: carbon-monoxide oxidation.

standardise the annotations and builds their corresponding GEM. These GEMs can later on be used in the two analysis involved in the pipeline: the assessment of metabolic capabilities using RetroPath2.0 [38] and the viability of community growth using MICOM [35]. The RetroPath2.0 analysis focuses on determining whether interesting metabolites, from an Astrobiological perspective, can be produced using the already present reactions in the community. MICOM is used in parallel to assess how this community can grow subject to this production constraint and to explore what can be the dominant species and any synergies or antagonisms that may arise. Finally, the results from both tools are combined to assess whether the production of the given compound subject to the community and environment constraints is feasible.

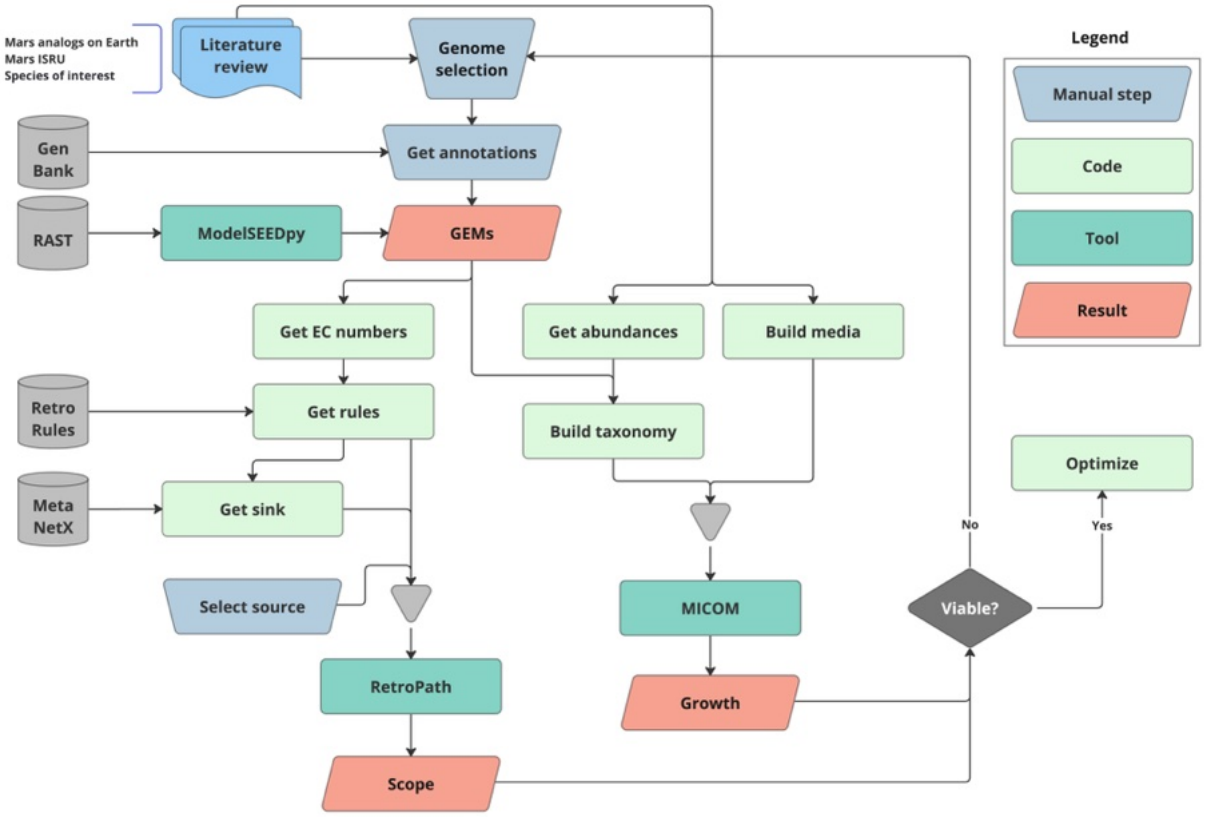


Figure 6: Flowchart followed for exploring the microbial communities in Río Tinto. The flow of goes from top to bottom. First, literature review yields a set of genomes for which annotations are retrieved. Then, GEMs are created from these annotations and are used by both RetroPath2.0 and MICOM for running the analyses. Finally, viability is assessed by inspecting the results from both tools, followed by optimization in case of promising outcomes or re-entry in a new iteration if required.

2 Objectives

Given the importance of Río Tinto as a Mars analog, we decided to build a data pipeline for exploring its potential in helping set up a hypothetical biofoundry on Mars:

- Integrate information for the species involved from different data sources.
- Create models of the different metabolisms involved in this environment.
- Assess whether those species can synthesize a compound of interest.
- Study microbial community growth under different conditions and the effect of target compound production on it.
- Re-purpose algorithms used in metabolic engineering, synthetic biology and microbiome analysis for Astrobiology research.
- Propose an agnostic data pipeline for studying analog environments on Earth.

3 Materials and Methods

3.1 Software

The pipeline depicted in Figure 6 has been implemented in Python [42] and can be found in this GitHub repository: guillecg/mars-biofoundry (see Figure 7). Several numerated Jupyter notebooks [43] guide the user from the data acquisition and wrangling to the creation of the metabolic models to finally launch RetroPath2.0 and MICOM. All plots used in this work are generated using these notebooks. Software development best practices such as DRY [44], KISS [45], YAGNI [46] and SOLID [47] have been followed throughout the code. Additionally, all functions have been documented using docstrings[4] following the NumPy style[5].

3.2 Genome selection

Of the total number of organisms in [1], we selected those with an annotated genome from which we could later on build a metabolic model using ModelSEEDpy[6] [40]. This requirement enormously reduces the number of

[4] https://peps.python.org/pep-0257/
[5] https://numpydoc.readthedocs.io/en/latest/format.html
[6] https://github.com/ModelSEED/ModelSEEDpy

species by an order of magnitude, leaving only ten out of hundreds (see Table 1). While the variability of potential metabolisms in the community is heavily compromised, our selection captures most of the critical pathways as detailed in [1] (see Figure 4). In fact, most of the pathways are represented with the exception of the carbon and sulfur cycles, for which there are no representatives for some steps (especially in methane-related pathways).

Species	Code	Database	ID	Protein annotation file	Gene count
Acidovorax BoFeN1	aci	NCBI	QOZT00000000.1	QOZT01.1.fsa_aa	3672
Brevundimonas sp. T2.26MG-97	bme	NCBI	NZ_UXHF01000001.1	NZ_UXHF01000001.1.faa	133
Desulfosporosinus meridiei DEEP	dmi	IMG	2721755100	CP003629.1.faa	4352
Pseudomonas sp. T2.31D-1	pse	ENA	CAJFAG010000000.1	CAJFAG01.1.fsa_aa	4428
Rhizobium sp. T2.30D-1.1	rhi1	NCBI	NZ_UEYP01000001.1	NZ_UEYP01000001.1.faa	756
Rhizobium sp. T2.26MG-112.2	rhi2	NCBI	NZ_UEYQ01000001.1	NZ_UEYQ01000001.1.faa	711
Rhodoplanes sp. T2.26MG-98	rho	NCBI	NZ_UWOC01000001.1	NZ_UWOC01000001.1.faa	19
Shewanella sp. T2.3D-1.1	shw	ENA	CACVBT0200000010	CACVBT03.1.fsa_aa	4068
Tessaracoccus lapidicaptus IPBSL-7	tel	IMG	2791354959	MBQD01P.1.fsa_aa	2736
Tessaracoccus sp. T2.5-30	tez	IMG	2751185744	CP019229.1.faa	2905

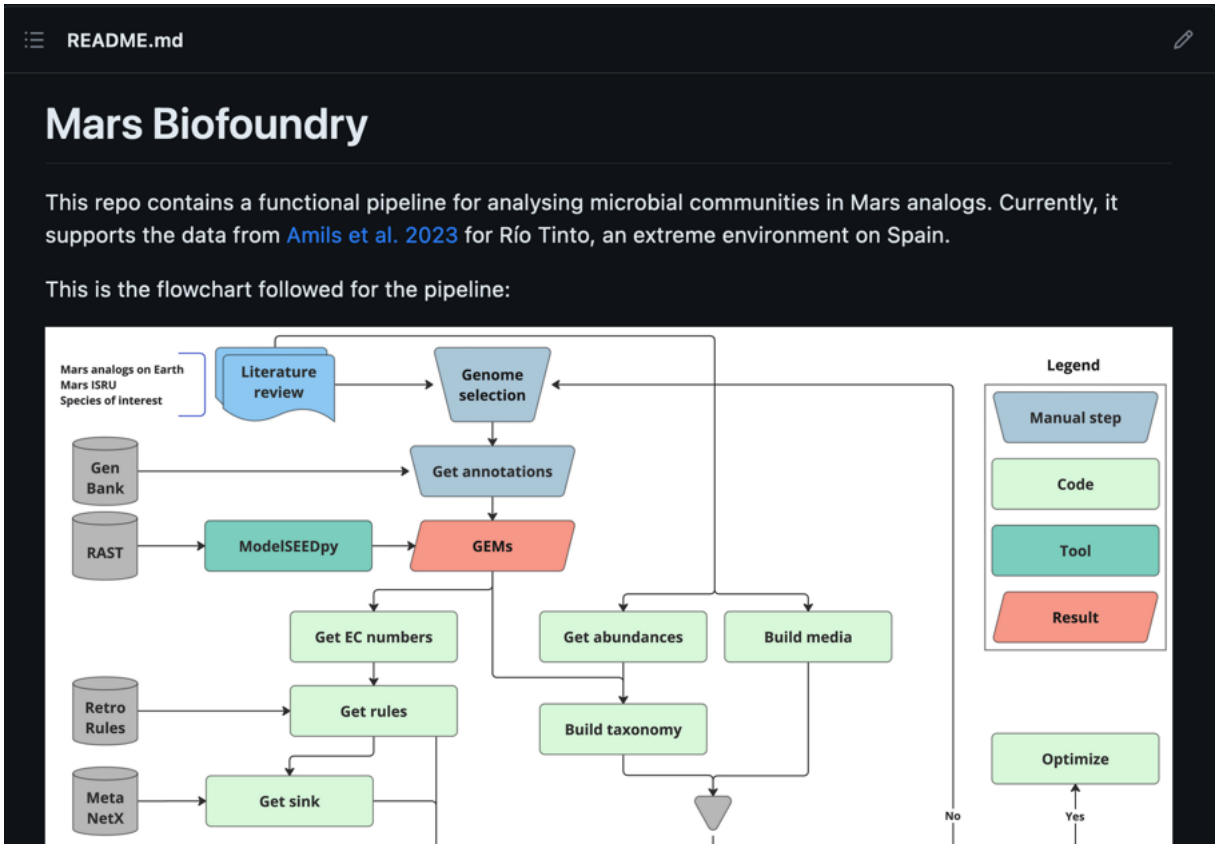
Table 1: Species analysed in this work. These species capture most of the cycles depicted in Figure 4 and can serve as a base community to later optimize it. The table also contains the exact file names as shown in their respective sources, in order to facilitate the reproducibility of the results.

3.3 Metabolic model reconstruction

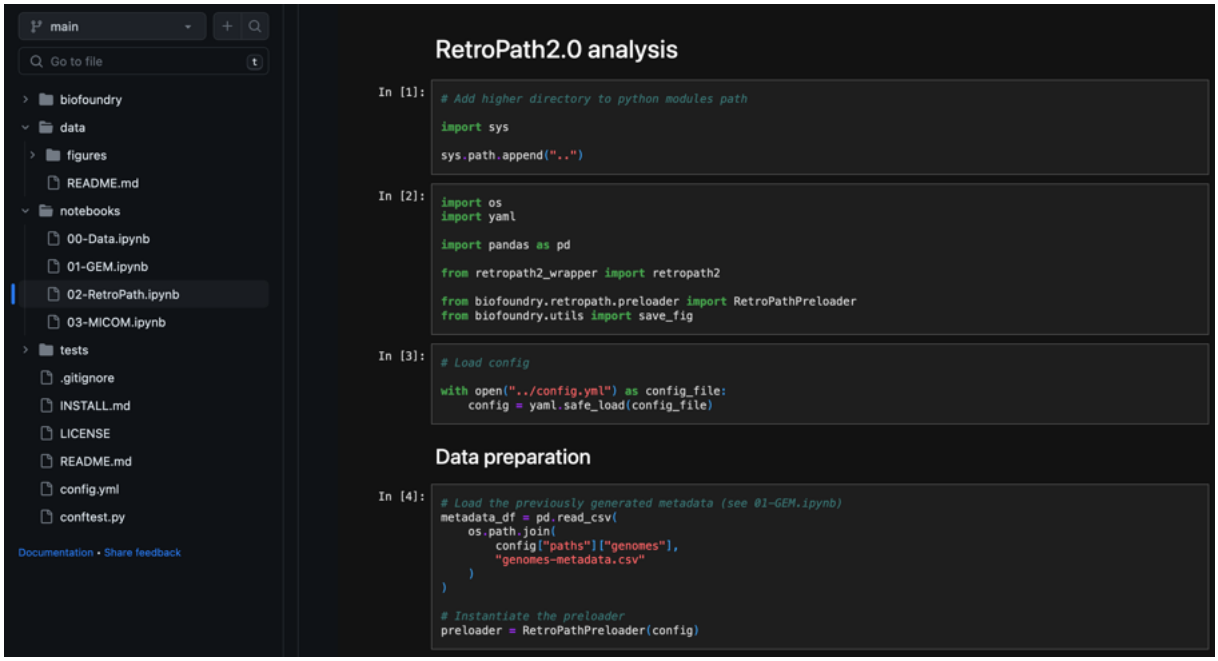
The process starts with the sequenced genome of a species of interest. Annotation of that genome is required in order to know which genes it contains, some of which will codify for a certain type of proteins known as enzymes. These proteins are responsible for catalyzing the biochemical reactions that compose the metabolism and their function is usually denoted by an Enzyme Commission (EC) or Transport Commission (TC) number [48, 49, 50]. GEMs integrate information from reactions and metabolites using a stoichiometric matrix, which defines the relationships between them [23, 33, 51] (see Figure 8). By identifying which enzymes each species contains, we can further elucidate which reactions can those organisms perform and, thus, the metabolic potential present in a given community. Table 1 details the database and the identifier from where each annotation was retrieved as well as the name of the file.

The creation of a metabolic model from an annotated genome has been automatised in several publications, such as ModelSEEDpy[7] [40], PyFBA [51] or CarveMe [52]. We decided to work with ModelSEEDpy because of its ongoing development and maintenance as well as its relative higher

[7] https://github.com/ModelSEED/ModelSEEDpy



(a) Repository front page.



(b) Jupyter notebook for launching RetroPath analysis.

Figure 7: GitHub repository containing the agnostic pipeline. The repository contains all the information (a) and related notebooks (b) guiding the user towards successfully launching the whole pipeline, from data ingestion to the generation of figures. It is available here: guillecg/mars-biofoundry.

tests coverage. Its ready-to-use functions also facilitate the integration with our pipeline. In order to better assess the quality of the created GEMs, a benchmark should be performed with all of the tools (see Section 6). Nevertheless, manual curation is still a critical issue in this step.

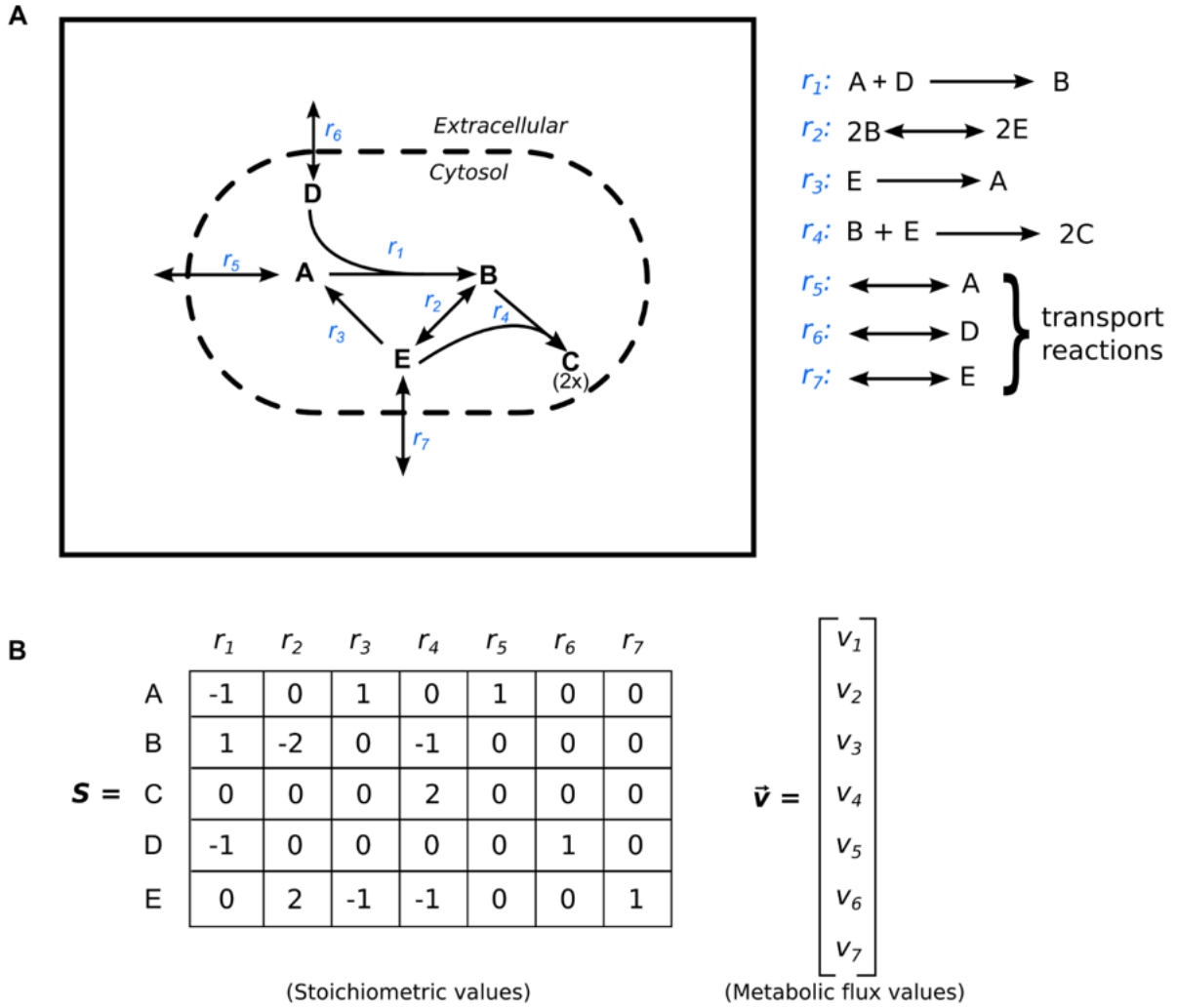
3.4 RetroPath: Assessment of metabolic capabilities

To evaluate the metabolic potential of the microbial community in the subsurface of Río Tinto, we explored the production of different compounds of interest, both from a perspective of a potential colony on Mars or because of their biotechnological value. Many tools are available that can predict or enumerate such production pathways [23], but we decided to use RetroPath v2.0 [38] because of its ability to retrosynthetically fill the gaps for synthetizing a target compound. Additionally, it supports compatibility with the extensive RetroRules database [53]. RetroPath is implemented in KNIME [54] and comes with a tutorial for easily running the examples provided and modify them for tailored purposes.

Retrosynthesis algorithms take as input a set of metabolites (organisms' compounds) and a set of target molecules (compounds of interest) and generate the metabolic networks required to link the former (usually denominated sink) to the latter (usually denominated source) [38] (see Figure 9). To perform such mapping, RetroPath requires the source, the sink and the reactions (rules) that build up the networks.

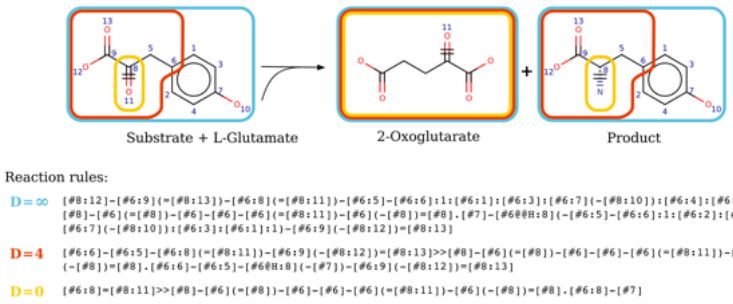
Rules selection. For extracting all rules from the GEMs, we first needed to map the reactions within the model to EC numbers. Querying RetroRules [53] with those identifiers returns the rules required to build the chassis of our community (see Figure 9a). It is worth mentioning that there are forward (naturally occurring), reverse and reversible (both directions) rules in the database. We selected only those belonging to the forward and reversible categories. A total number of 11589 EC numbers and 42664 rules were extracted for the community.

Sink selection. To check whether the source can be produced, a set of compounds, both present in the microorganisms and in the medium, must be provided. This set of nutrients and intermediary metabolites establishes the baseline from which the species can grow and produce the target. Since the compounds involved in each rule are defined, we decided to extract these molecules for building the sink: substrates both for forward and reversible rules, while only products for reversible ones. Ad-

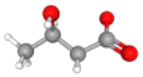


	r_1	r_2	r_3	r_4	r_5	r_6	r_7
A	-1	0	1	0	1	0	0
B	1	-2	0	-1	0	0	0
C	0	0	0	2	0	0	0
D	-1	0	0	0	0	1	0
E	0	2	-1	-1	0	0	1

Figure 8: Conversion of a simple metabolic network into a mathematical format. **(A)** Example of a bacterial metabolic model displaying two compartments separated by a dashed boundary (extracellular and cytoplasm), seven reactions labeled in blue text (four intracellular and three transporters), and five compounds. **(B)** The stoichiometric matrix S with corresponding stoichiometric coefficients, and the flux vector v. Each matrix-cell represents the number of compound molecules required for the particular reaction. The integer sign denotes the compound as a reactant (negative value) or as a product (positive value). A zero means the compound is not involved in the reaction. Reversible reactions are typically present in the matrix in one direction. In the instance that a reaction is reversed in the solution, the metabolic flux value for the corresponding reaction will be negative, thus indicating a switch in directionality. Figure and caption taken from [51] (CC BY 4.0).



(a) Example rule: EC 2.6.1.5. Figure adapted from [53] (CC BY 4.0).



(b) Example source: 3-hydroxybutyrate. [60].

Figure 9: Examples of a rule (a) and a source (b) given as input to RetroPath2.0. (a) Rules are extracted from RetroRules [53] whereas sources are user-defined and often manually selected. Top: 2D compound representation of the reaction with different atom diameters highlighted in yellow (D=0), red (D=4) and blue (D=inf). The lower the diameter, the higher the specificity and vice-versa. Bottom: SMIRKS[8]codifications of the same reaction, which allows for more generic representations of such chemical transformations. **(b)** 3-hydroxybutyrate is the monomer of PHB, a bioplastic with potential for helping in establishing a biofoundry.

ditionally, the identifiers in RetroRules come from MetaNetX [55], so we further queried this database to pull the information for each compound, including its International Chemical Identifier (InChI), a unique identifier for each molecule [56, 57]. A total number of 1182 unique metabolites were used as the sink for the community.

Source selection. The source or target compound must be manually selected by the user, and we decided to start with 3-hydroxybutyrate, the building block of the polymer poly-3-hydroxybutyrate (PHB) (see Figure 9b). Given that plastics will make up the majority of high-turnover items and can also account for contingencies [11], PHB has the potential to fill this gap as a bioplastic and because is naturally produced by some microorganisms as a carbon and energy reservoir [58]. Additionally, a set of 476 compounds with biotechnological interest [59] was also screened to check for any other targets that the community may produce.

RetroPath launch. RetroPath was launched using KNIME v4.6.4 [54]. The workflow was downloaded as specified in [38] and the parameters set to their default values with the exception of maximum diameter (16), minimum rule diameter (6) and pathway length (10). Launch was performed programmatically using RetroPath2-wrapper[9].

[8] www.daylight.com/dayhtml/doc/theory/theory.smirks.html
[9] https://github.com/brsynth/RetroPath2-wrapper

3.5 MICOM: Community analysis

Since no organism thrives in isolation, we wanted to holistically explore how the community could growth and how different species could interact between them, by means of cooperation or antagonism. The package MICOM [35] allows to study the behaviour of a given community by exploring different tradeoffs among the members of it.

Retrieval of abundances. MICOM requires the abundances of the different species in the samples analysed to account for the variability between samples and species. For our study, abundances were extracted from the supplementary materials in [1] (see Figure 10). Concretely, datasets S4 and S5 were used, containing data from Illumina and Roche 454 sequencing technologies, respectively. The reads used correspond to the filtered species, removing those that could potentially imply a contamination in the samples. It is worth mentioning that only *pse* was found in Roche samples.

As Figure 10 depicts, there are many gaps in the abundances of the selected species for this study, that is, the species with reconstructed genome. Abundances for hundreds of more species can be found at the supplementary materials, but there are no corresponding annotated genomes that can be used in the proposed pipeline. This situation strongly limits the applicability of the current pipeline for exploring the microbial community as is in its natural environment, forcing us to focus on the metabolic engineering aspect of it. Therefore, several experiments are proposed to analyse its biotechnological potential.

An initial analysis focuses on what would be the differences among the community members when their abundances are set equally:

- 1.1. Objective: community growth (aerobic)[10]
- 1.2. Objective: community growth (anaerobic)[11]
- 1.3. Objective: production of metabolite of interest (3-hydroxybutyrate)
- 1.4. Objective: community growth (aerobic) + production of metabolite of interest (3-hydroxybutyrate)
- 1.5. Objective: community growth (anaerobic) + production of metabolite of interest (3-hydroxybutyrate)

[10] Aerobic: presence of oxygen.
[11] Anaerobic: absence of oxygen.

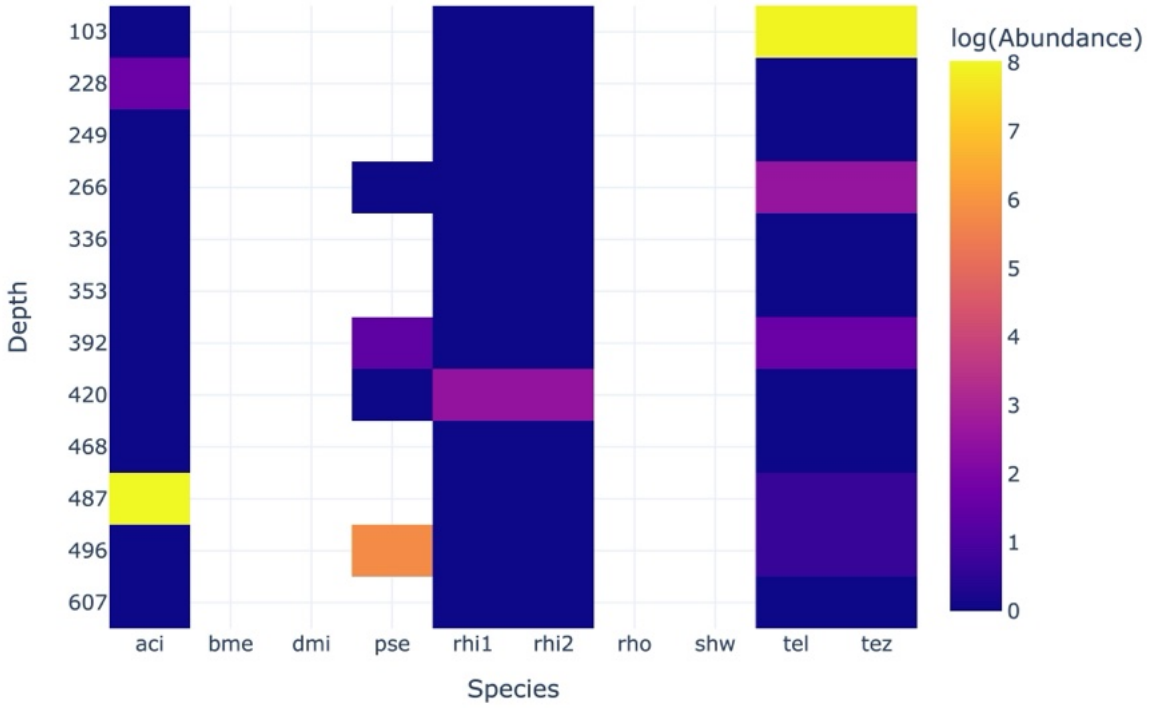


Figure 10: Abundances extracted from supplementary information in [1]. The high sparsity of the data is depicted, highlighting four species for which there is no abundance information (*bme*, *dmi*, *rho* and *shw*). Abundances are also imbalanced, as can be noted in the enormous differences in orders of magnitude for some species (*aci*, *tel* and *tez*). Gaps in *pse* column are caused by that species being only present for Roche samples at those four different depths. Roche data does not contain information for the rest of the depths present in Illumina. Note that the values are the logarithm of raw abundances.

Taxonomy definition. The next step involves the creation of a taxonomy, an artifact that contains both the species metabolic models and the abundances specified for each one. Using this taxonomy, MICOM is able to predict the growth of the community as a whole, indicating which are the species with more or less growth given such conditions.

Medium reconstruction. Information from the following supplementary materials in [1] was used for medium reconstruction:

- Dataset S2 - ICP-MS elemental analysis of core samples (ppm)
- Dataset S3 - Ionic chromatography of BH10 soluble organic and inorganic anions (ppm)
- Table S1 - Soluble cations (ppm)

- Table S7 - Occluded gases and natural activities at different depths (ppm)

The data were merged into a single file according to their measurement depth, which can later on be used for matching the abundances of species at different depths.

4 Results

4.1 Reconstruction of metabolic models

Six out of ten metabolic models show an acceptable ratio between the amount of annotated genes and the number of metabolites and reactions in their respective GEMs (see Figure 11). For the remaining four, the amount of annotated genes is quite restrictive, being just 133 for *bme* in the input data and reaching the lowest in *rho*, for which just 19 genes are present in its annotated genome. This situation is further aggravated by the automatic curation performed by ModelSEEDpy when generating the metabolic model. Regarding the previous examples, 28 and 4 genes are present in the GEMs corresponding to *bme* and *rho*, respectively. Special consideration should be taken when handling these restricted models. It is worth noting the baseline number of reactions and metabolites present in both organisms correspond to the core model that ModelSEEDpy uses as template[12].

4.2 RetroPath

4.2.1 Production of 3-hydroxybutyrate.

The microbial community presented 3-hydroxybutyrate as one of its metabolites, being present in *aci* (MNXM1104965), *rhi1* (MNXM1104967) and *tel* (MNXM1104965 and MNXM1104966). Interestingly, both enantiomers (R: MNXM1104965, S: MNXM1104966) as well as the compound without specified isomerism (MNXM1104967) were found.

4.2.2 Production of compounds of interest.

The set of 476 compounds extracted from [59] was screened using RetroPath [38] to find any other compound with biotechnological interest that could be produced by the community. Figure 12 illustrates how many of these

[12] https://github.com/ModelSEED/ModelSEEDpy/blob/dev/modelseedpy/core/msbuilder.py

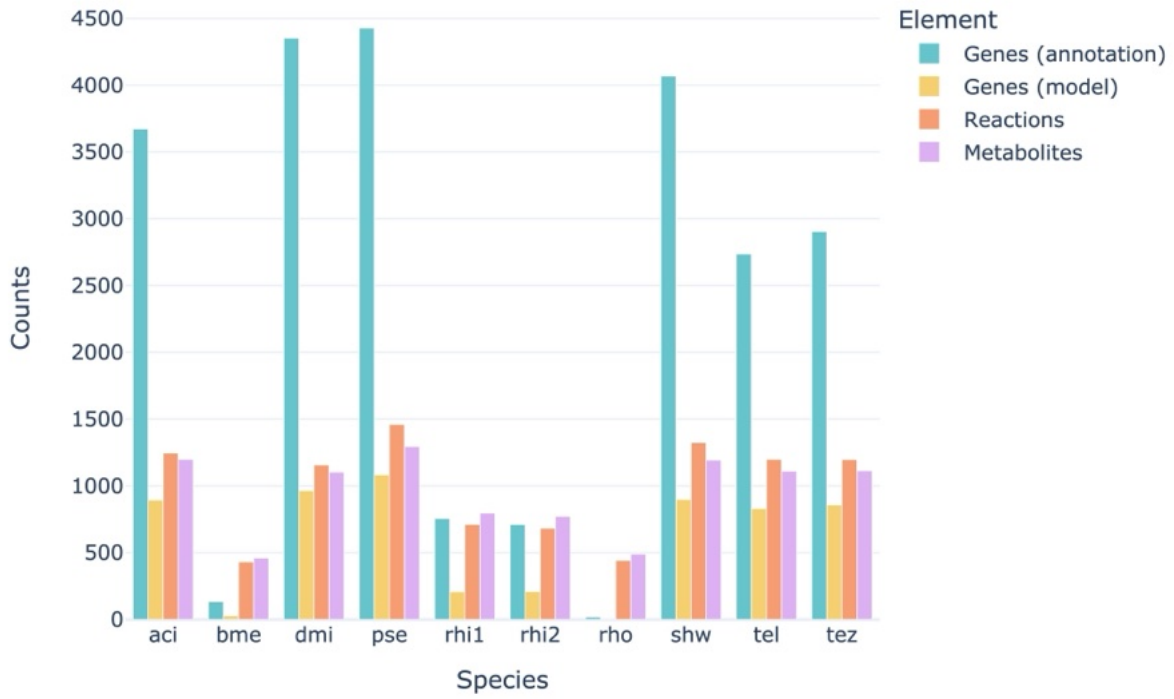


Figure 11: Summary of metabolic models produced by ModelSEEDpy. Note the difference in counts between the genes in the input annotation files and the final number of genes in the model. This is caused by the presence of a high proportion of genes without a clear annotation, thus yielding hypothetical functions that are eliminated by ModelSEEDpy. Additionally, four models contain either a low number of genes compared to other ones (*rhi1* and *rhi2*) or an almost inexistent quantity (*bme* and *rho*). The presence of a baseline number of reactions and, therefore, metabolites is due to the use of a core metabolism in the creation process.

targets can be produced. Attending to the different categories, *Source in sink* accounts for those metabolites that are already found in the community and, thus, susceptible to be directly optimized for industrial production. Compounds in the category *Scope* are also susceptible to be produced by adding rules present in the community, although lacking in the species individually. This category may help identify compounds that are shared in the environment and cooperatively used by different species, given that its production is achieved thanks to the combination of rules present in different species. The rest of categories can be considered run errors.

Focusing on the metabolites for which a scope was found, Figure 13 highlights the strong skew towards phenylacetate and tyrosine compound classes that can potentially be synthesized from the community. It is worth not-

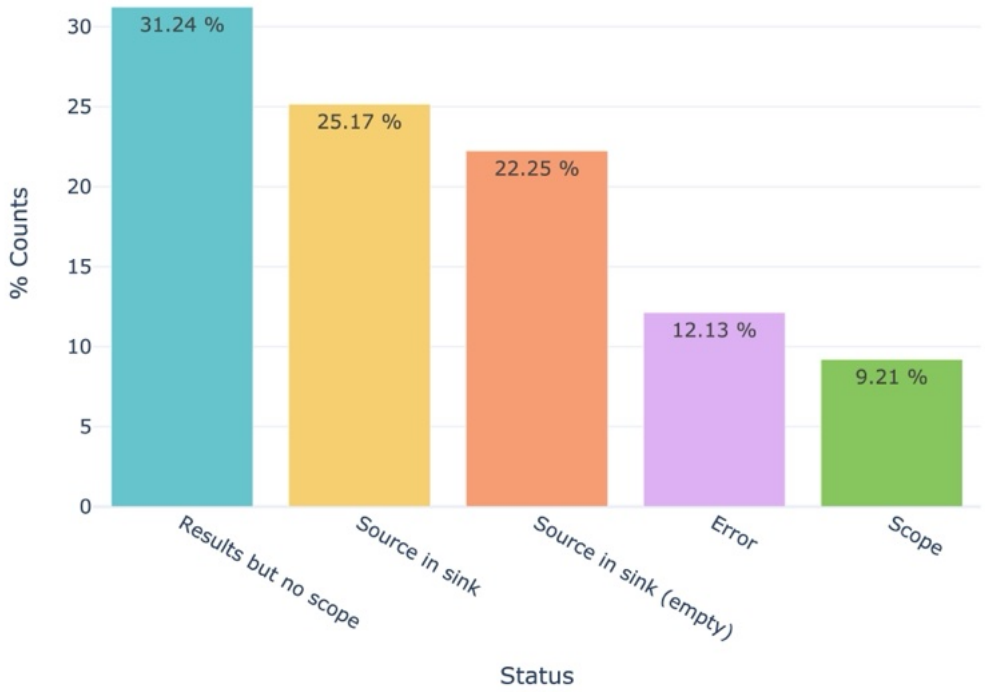


Figure 12: Summary of RetroPath2.0 results for the microbial community in Río Tinto and the metabolites with biotechnological interest. Of all categories, the ones that cannot be considered errors or failures to find a solution are *Source in sink* and *Scope*. Overall, more than a third of the interesting metabolites are either already present or could potentially be produced by the community.

ing that all statuses show this skew towards these two classes, which could imply a bias in the original classification used. Energy-related classes can also be found in this subset of compounds, such as amino acid biosynthesis (valine, leucine and isoleucine) and degradation (lysine, tyrosine) as well as sugar metabolism (fructose and mannose). Interestingly, classes related to photosynthesis are present, as well as several categories involving benzene derivatives and aromatic compounds, such as benzoate, nitrophenol, bisphenol, phenanthrene, nitrotoluene and xylene.

4.3 MICOM

4.3.1 Medium reconstruction

Due to the difficulty in mapping from concentration in the environment to real fluxes in the cell, we extrapolated concentrations using the limiting metabolites. MICOM and COBRApy [64] (on top of which MICOM is based) offer a functionality for finding the minimal medium, that is, the

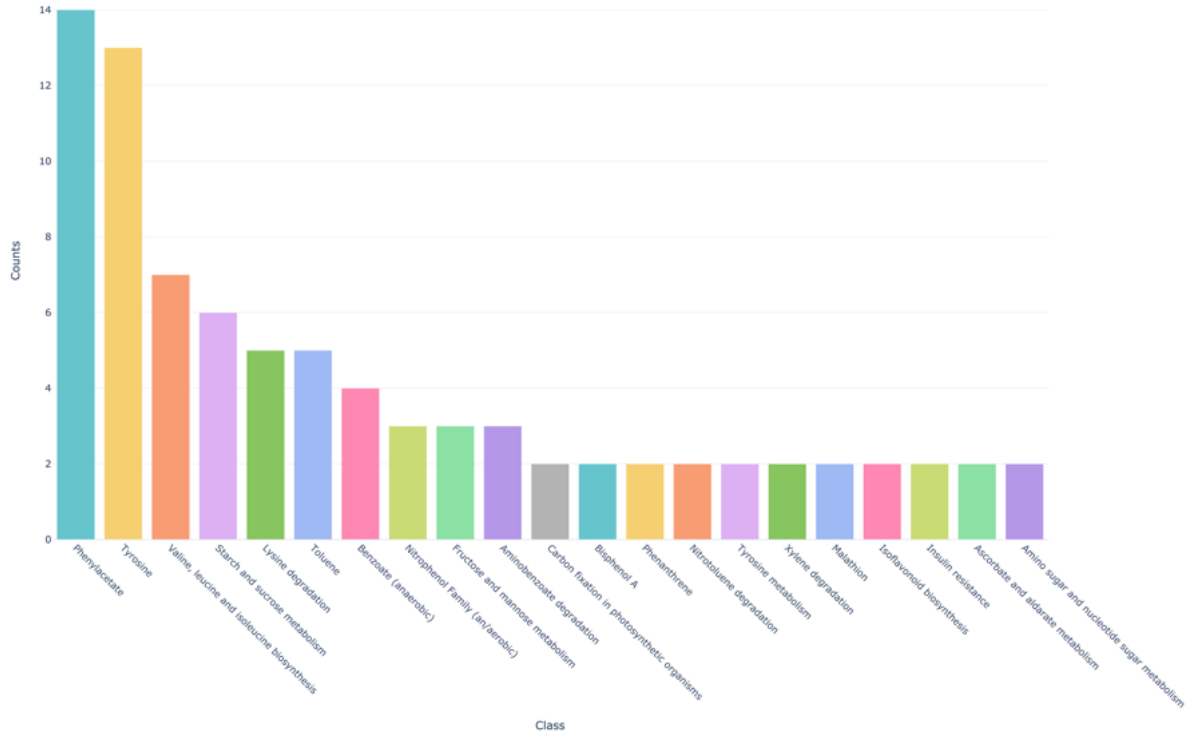


Figure 13: Distribution of compound class occurrences for the metabolites with found scope. Counts have been capped at higher or equal than 2 due to the high skewness of the distribution. Compound classes were generated by Lorena Martínez España and Hèctor Martín Làzaro using iFragment [61] with KEGG [62] and EnviPath [63] databases.

set of minimal metabolites and their fluxes that the species (community in our case) requires for growing (see Figure 14). By finding the metabolites present in the medium from [1] in this minimal medium, we can extrapolate the fluxes to the rest of compounds for which no information is provided. In our case, nitrate (NO_3^-, *EX_cpd00209=no3_m* in Figure 14) was the pivotal metabolite for matching compound concentration and fluxes.

4.3.2 Community growth

Figure 15 shows the results obtained for each case study in section 3.5. Given that the abundance data was really limited for most of the annotated genomes (see Figure 10), we decided to conduct the experiments assuming that all species were present in the same proportion. Species with the lowest quality of GEMs (*bme*, *rhi1*, *rhi2* and *rho*) show almost no differences between experiments. However, a potential dependency on molecular oxygen for growth is shown for the rest of species and the community as a whole, as illustrated by the lower values in anaerobic versus

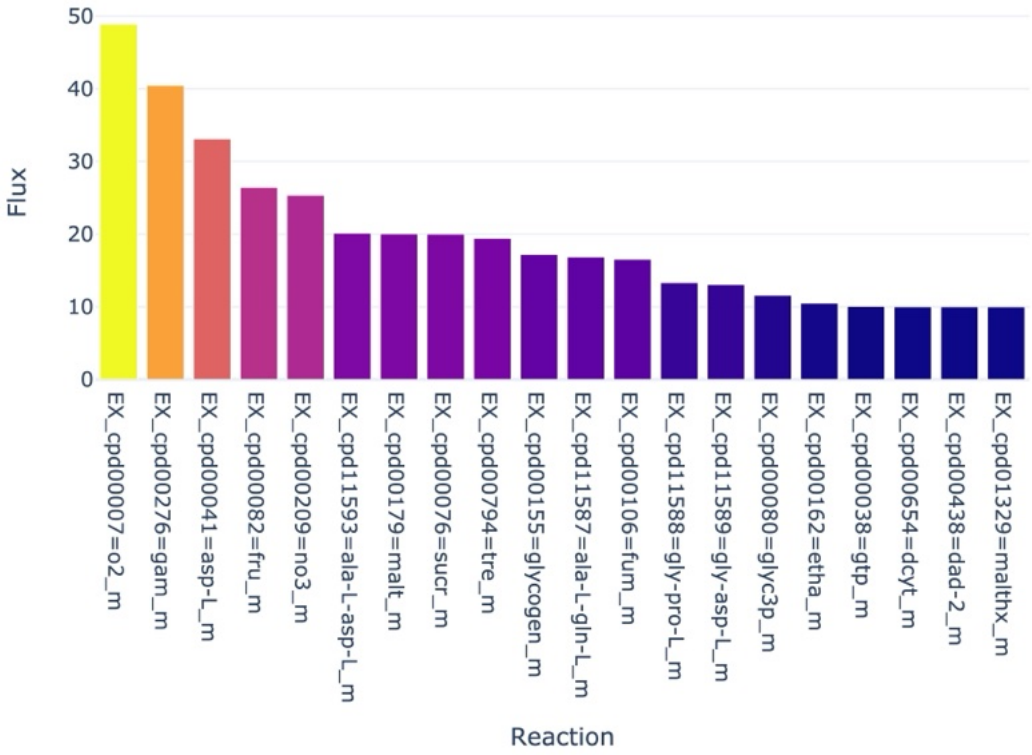
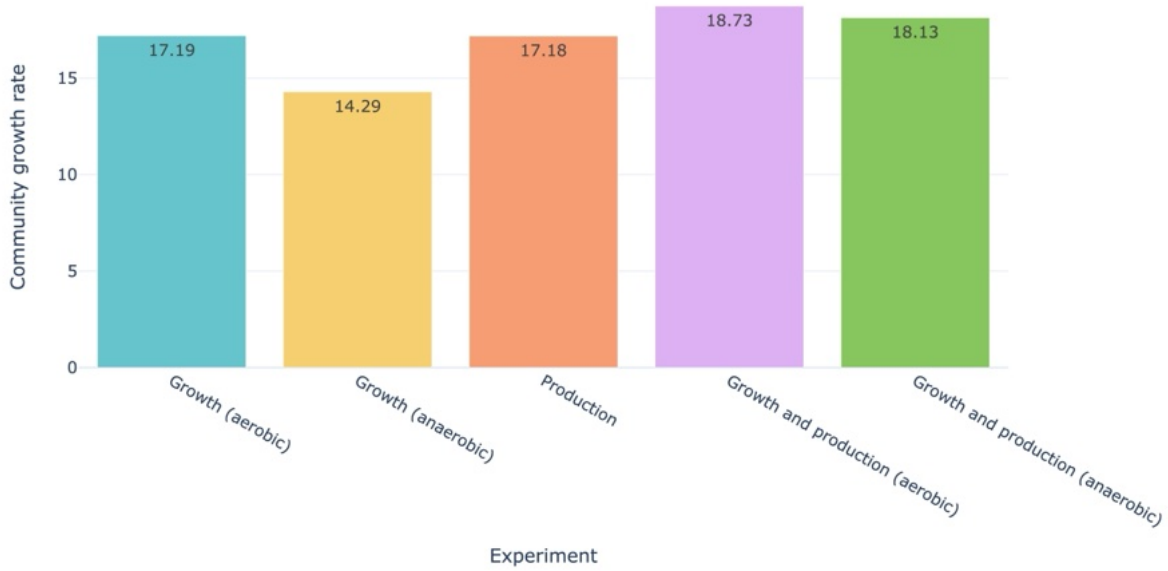
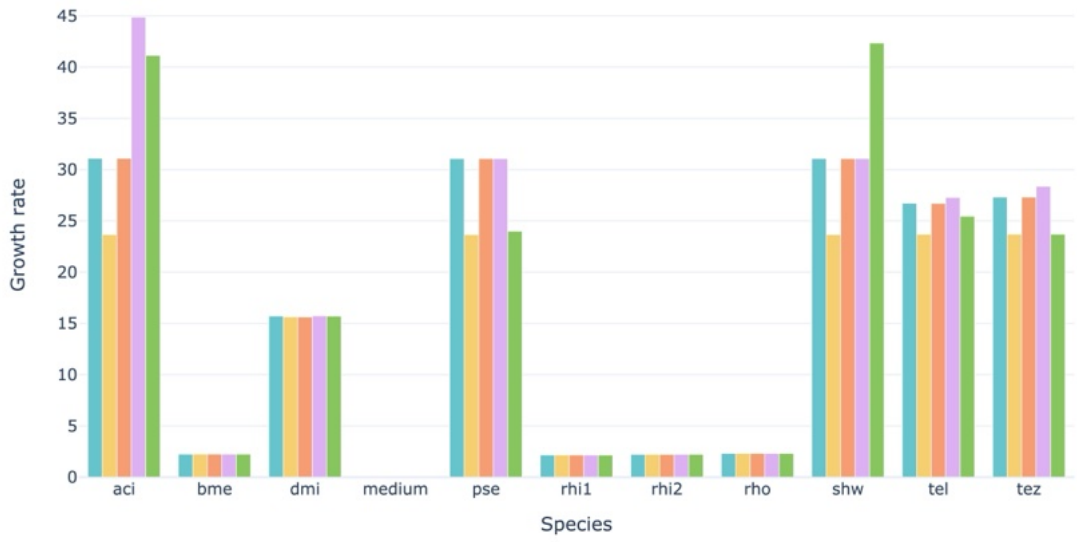


Figure 14: Limiting metabolites for the community built with equal abundances. The identifiers are created as following: *EX* for exchange with the environment, ModelSEED ID, compound abbreviation and *m* for medium. Sugars (glucosamine, *gam*; fructose, *fru*; maltose, *malt*; sucrose, *sucr*; trehalose, *tre*; glycogen) and amino acids (aspartic acid, *asp*; alanine and aspartic acid dipeptide, *ala-L-asp-L*) dominate the upper range of fluxes. Interestingly, molecular oxygen (O_2) is the top limiting metabolite, while nitrate (NO_3^-) will allow the extrapolation of compound concentrations with fluxes.

aerobic conditions as well as the presence of oxygen as the top limiting metabolite (see Figure 14). The exception is *shw*, which shows a higher value for growth and production under anaerobic conditions, in concordance with its anaerobic nature [65]. Overall, *aci*, *pse* and *shw* dominated the growth within the community, with *tel* and *tez* following them to a lower extent. On the other hand, *dmi* behaved in a similar manner to species with lower quality GEMs (no differences across experiments), but with a higher growth rate.



(a) Community growth.



(b) Species growth.

Figure 15: Community (a) and species (b) growth in the different experiments. Color and categories in (a) correspond to the same in (b). Unsurprisingly, species with the lowest quality of GEMs (*bme*, *rhi1*, *rhi2* and *rho*) show minimal differences across experiments. Those differences, however, are accentuated between aerobic and anaerobic conditions, suggesting a dependency on molecular oxygen.

5 Conclusions

An agnostic data pipeline has been implemented for assessing the production of compounds of interest and the community growth in extreme environments, all from an Astrobiological perspective. This pipeline is available in a public GitHub repository[13] and will be further developed on, with the objective of publishing a Python package in the future. Additionally, Río Tinto's microbial community was analysed using the guidelines provided in the repository and the results are presented in this work.

Recapitulating the objectives of this study:

- We successfully integrated information from different databases, such as GenBank [39], ModelSEED [40], RetroRules [53] and MetaNetX [55] as well as the abundant supplementary information provided in [1]. An important aspect of this integration is the preliminary literature review, to determine the gaps and critical steps in ISRU.
- Apart from enriching the data and facilitating the reconstruction of metabolic models, this integration of different data sources also allowed the study of the microbial community from reductionist and holistic perspectives, such as the production of compounds of interest or the community growth, respectively.
- Precisely, the pipeline could not have been possible without the repurposing of tools used in metabolic engineering, synthetic biology and microbiome analysis, for example to create the metabolic models or GEMs (ModelSEEDpy[14] [40]), to assess the production of interesting compounds (RetroPath [38]) and to simulate the microbial community under specific conditions for ISRU (MICOM [35]).
- Regarding the implemented pipeline, a series of Jupyter notebooks [43] guide the user from the data acquisition and preprocessing to the creation of GEMs and launch of RetroPath and MICOM. The decoupling of these steps from the preprocessing of the input data increases the flexibility of the proposed pipeline, as well as facilitates its usability for the users. They only need to adapt their input data to the format required by this workflow in order to be able to launch it. Additionally, the repository contains guides for the installation of de-

[13] https://github.com/guillecg/mars-biofoundry
[14] https://github.com/ModelSEED/ModelSEEDpy

pendencies[15] and the retrieval of necessary data from public sources[16].

Notwithstanding, several limitations to the current analysis must be noted:

- **Genome annotations.** The low quantity (ten out of hundreds) as well as quality of the annotated genomes lowers the amount of cycles and metabolic functions that can be captured.
- **Metabolic models.** Low quality in annotated genomes further cascades into the reconstruction of metabolic models, which strongly limits the utility of the results obtained.
- **Medium reconstruction.** Although data coming from several measurements in [1] was successfully integrated, their usage as medium in MICOM is seriously limited by the limiting metabolites that can be found, from which to extrapolate the fluxes (see Section 4.3.1).
- **MICOM.** Reduced number of species and abundances make difficult the analysis, again reducing the useful information that can be extracted from it.

The true utility of this framework will come after successfully applying it to a new environment. By testing the pipeline on unseen data, we will be able to validate whether the current architecture fits the potential issues with the format and files. [30] could be a good starting point, given that it analyses the water and sediments of Río Tinto. All in all, this agnostic data pipeline lays the scaffold for further integration of data sources and addition of tools, such as DeepFRI [66] (see Section 6), which could expand its potential for exploring these captivating and mysterious environments.

6 Future work

Add additional genome annotations. One of the most limiting characteristics of the data is the low amount of annotated genomes present. Increasing the number by looking at phylogenetically close species can prove useful, at the cost of robustness in the analysis, since some species may end up having dissimilar metabolic functions.

Improve genome annotations. Quality genome annotations are a limiting factor in our pipeline. Many annotated genes are predicted as hypothetical or directly lack any function. Therefore, automatically annotating

[15] https://github.com/guillecg/mars-biofoundry/blob/main/INSTALL.md
[16] https://github.com/guillecg/mars-biofoundry/tree/main/data

genomes can help retrieve more functions from the input data, thus enriching the metabolic models at the cost of lower quality. One of such tools that can be adapted to work in this framework is DeepFRI [66]. By leveraging both sequence and structure, it is capable of predicting the function (e.g. EC number [48, 49]) of a given input sequence. A first attempt at adapting this tool to the pipeline was performed, however halted by the lack of specificity of DeepFRI's predictions. EC numbers are more specific the more digits are provided (i.e. an EC number with four digits is more specific than one with just two). Unfortunately, DeepFRI yields EC numbers without the last position, yielding a wide range of enzymes for a single EC number prediction, enzymes that in most cases are strongly substrate-specific and cannot catalyze the reactions of other entities under the same EC. For example, 97 entries are found for EC 1.13.11.-[17], which belong to oxydoreductases involving molecular oxygen on a very broad range of substrates, some of them very dissimilar between them (e.g. tryptophan (EC 1.13.11.11) and linoleate (EC 1.13.11.12)), suggesting a high substrate specificity. Therefore, all these predicted EC numbers cannot be added as is to the genome annotations and further curation is required.

Benchmark and scale up. Several tools in the literature can build a GEM from the annotated genome [37, 51, 52]. For example, metaGEM [37] leverages CarveMe [52] and SMETANA[18] [36] among other tools for building the community metabolic models from raw metagenomic data, for which we didn't have access to in this case study. Therefore, performing a benchmark will allow their comparison and select the most suited for our setting. Additionally, different algorithms can be used to fill gaps between metabolites in the models, a process called gapfilling [34]. Scaling the analysis to the combination of model creation methodologies plus gapfilling algorithms could shed some light on the quantitative improvements between tools. Additionally, some packages, such as COBREXA.jl [67], may help scaling up the analysis to higher orders of magnitude with regards to the species considered.

Manually curate results. Using RetroPath2.0 [38] for assessing the capability to produce a certain compound is useful, but a biofoundry will require further results in order to adapt to different target compounds or conditions. Exploring and cataloging the set of reactions that can be useful to such ends will provide essential in a proof-of-concept. RetroRules [53]

[17] https://enzyme.expasy.org/EC/1.13.11.-
[18] https://github.com/cdanielmachado/smetana/tree/master

can be used as a source of unseen reactions in the community, but manual curation will likely be needed. Additionally, GEMs and, by extension, results from MICOM strongly depend on the quality of the annotations. Although extremely time consuming (from 3 months to 1 year according to [33]) and out of the scope of this work, manual curation of GEMs is required in order to further dissect and solve the different issues that have been arising along the pipeline. Tools such as MEMOTE [68] could help in this endeavour.

Explore synergies and antagonism. MICOM can use the change in correlation after species knockouts to check for any synergies or antagonism. This analysis will prove extremely useful for newly, less studied environments and for establishing groups of species that will likely be cooperating in different settings. Deviations from these measurements can also be useful to distinguish between strong and soft relationships. However, the limited number of species involved in our work limits the utility of this kind of plots, requiring more species to really capture groups of organisms that work together.

References

[1] R. Amils, C. Escudero, M. Oggerin, F. Puente Sánchez, A. Arce Rodríguez, D. Fernández Remolar, N. Rodríguez, M. García Villadangos, J. L. Sanz, C. Briones, M. Sánchez-Román, F. Gómez, T. Leandro, M. Moreno-Paz, O. Prieto-Ballesteros, A. Molina, F. Tornos, I. Sánchez-Andrea, K. Timmis, D. H. Pieper, and V. Parro, "Coupled C, H, N, S and Fe biogeochemical cycles operating in the continental deep subsurface of the Iberian Pyrite Belt," *Environmental Microbiology*, vol. 25, no. 2, pp. 428–453, Feb. 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/1462-2920.16291

[2] NASA, "Artemis." [Online]. Available: https://www.nasa.gov/specials/artemis/

[3] M. Smith, D. Craig, N. Herrmann, E. Mahoney, J. Krezel, N. McIntyre, and K. Goodliff, "The Artemis Program: An Overview of NASA's Activities to Return Humans to the Moon," in *2020 IEEE Aerospace Conference*. Big Sky, MT, USA: IEEE, Mar. 2020, pp. 1–10. [Online]. Available: https://ieeexplore.ieee.org/document/9172323/

[4] NASA, "Overview: In-Situ Resource Utilization." [Online]. Available: https://www.nasa.gov/isru/overview

[5] B. G. Drake, S. J. Hoffman, and D. W. Beaty, "Human exploration of Mars, Design Reference Architecture 5.0," in *2010 IEEE Aerospace Conference*. Big Sky, MT, USA: IEEE, Mar. 2010, pp. 1–24. [Online]. Available: http://ieeexplore.ieee.org/document/5446736/

[6] A. J. Berliner, I. Lipsky, D. Ho, J. M. Hilzinger, G. Vengerova, G. Makrygiorgos, M. J. McNulty, K. Yates, N. J. H. Averesch, C. S. Cockell, T. Wallentine, L. C. Seefeldt, C. S. Criddle, S. Nandi, K. A. McDonald, A. A. Menezes, A. Mesbah, and A. P. Arkin, "Space bioprocess engineering on the horizon," *Communications Engineering*, vol. 1, no. 1, p. 13, Jun. 2022. [Online]. Available: https://www.nature.com/articles/s44172-022-00012-9

[7] J. Cilliers, K. Hadler, and J. Rasera, "Toward the utilisation of resources in space: Knowledge gaps, open questions, and priorities," *npj Microgravity*, vol. 9, no. 1, p. 22, Mar. 2023. [Online]. Available: https://www.nature.com/articles/s41526-023-00274-3

[8] K. Sacksteder and G. Sanders, "In-Situ Resource Utilization for Lunar and Mars Exploration," in *45th AIAA Aerospace Sciences Meeting and Exhibit*. Reno, Nevada: American Institute of Aeronautics and Astronautics, Jan. 2007. [Online]. Available: https://arc.aiaa.org/doi/10.2514/6.2007-345

[9] NASA, "The Recent Large Reduction in Space Launch Cost." [Online]. Available: https://web.archive.org/web/20210801024432/https://ntrs.nasa.gov/citations/20200001093

[10] CSIS, "Space Launch to Low Earth Orbit: How Much Does It Cost?" [Online]. Available: https://aerospace.csis.org/data/space-launch-to-low-earth-orbit-how-much-does-it-cost/

[11] A. J. Berliner, J. M. Hilzinger, A. J. Abel, M. J. McNulty, G. Makrygiorgos, N. J. H. Averesch, S. Sen Gupta, A. Benvenuti, D. F. Caddell, S. Cestellos-Blanco, A. Doloman, S. Friedline, D. Ho, W. Gu, A. Hill, P. Kusuma, I. Lipsky, M. Mirkovic, J. Luis Meraz, V. Pane, K. B. Sander, F. Shi, J. M. Skerker, A. Styer, K. Valgardson, K. Wetmore, S.-G. Woo, Y. Xiong, K. Yates, C. Zhang, S. Zhen, B. Bugbee, D. S. Clark, D. Coleman-Derr, A. Mesbah, S. Nandi, R. M. Waymouth, P. Yang, C. S. Criddle,

K. A. McDonald, L. C. Seefeldt, A. A. Menezes, and A. P. Arkin, "Towards a Biomanufactory on Mars," *Frontiers in Astronomy and Space Sciences*, vol. 8, p. 711550, Jul. 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fspas.2021.711550/full

[12] COSPAR, "Panel on Planetary Protection (PPP)," Jul. 2023. [Online]. Available: https://cosparhq.cnes.fr/scientific-structure/panels/panel-on-planetary-protection-ppp/

[13] M. Rucker and S. Hoffman, "Explore Moon to Mars." [Online]. Available: https://ntrs.nasa.gov/api/citations/20205010563/downloads/COSPAR%20PP%20workshop%20Dec2020%20v6.pdf

[14] G. Sanders and J. Kleinhenz, "In Situ Resource Utilization (ISRU) Envisioned Future Priorities," May 2022. [Online]. Available: https://ntrs.nasa.gov/api/citations/20220004617/downloads/LIVE-ISRU%20Envisioned%20Future%20Priorities-Lux_SandersV3.pdf

[15] ESRIC, "ESRIC." [Online]. Available: https://www.esric.lu/

[16] R. Santomartino, N. J. H. Averesch, M. Bhuiyan, C. S. Cockell, J. Colangelo, Y. Gumulya, B. Lehner, I. Lopez-Ayala, S. McMahon, A. Mohanty, S. R. Santa Maria, C. Urbaniak, R. Volger, J. Yang, and L. Zea, "Toward sustainable space exploration: A roadmap for harnessing the power of microorganisms," *Nature Communications*, vol. 14, no. 1, p. 1391, Mar. 2023. [Online]. Available: https://www.nature.com/articles/s41467-023-37070-2

[17] M. Hecht, J. Hoffman, D. Rapp, J. McClean, J. SooHoo, R. Schaefer, A. Aboobaker, J. Mellstrom, J. Hartvigsen, F. Meyen, E. Hinterman, G. Voecks, A. Liu, M. Nasr, J. Lewis, J. Johnson, C. Guernsey, J. Swoboda, C. Eckert, C. Alcalde, M. Poirier, P. Khopkar, S. Elangovan, M. Madsen, P. Smith, C. Graves, G. Sanders, K. Araghi, M. de la Torre Juarez, D. Larsen, J. Agui, A. Burns, K. Lackner, R. Nielsen, T. Pike, B. Tata, K. Wilson, T. Brown, T. Disarro, R. Morris, R. Schaefer, R. Steinkraus, R. Surampudi, T. Werne, and A. Ponce, "Mars Oxygen ISRU Experiment (MOXIE)," *Space Science Reviews*, vol. 217, no. 1, p. 9, Feb. 2021. [Online]. Available: http://link.springer.com/10.1007/s11214-020-00782-8

[18] P. Gilster and A. Tolley, "MaRMIE: The Martian Regolith Microbiome Inoculation Experiment," Jan. 2023. [Online]. Available: MaRMIE:TheMartianRegolithMicrobiomeInoculationExperiment

[19] ESA, "MELiSSA life support project, an innovation network in support to space exploration." [Online]. Available: https://www.esa.int/Enabling_Support/Space_Engineering_Technology/MELiSSA_life_support_project_an_innovation_network_in_support_to_space_exploration

[20] NASA, "Veggie." [Online]. Available: https://www.nasa.gov/sites/default/files/atoms/files/veggie_fact_sheet_508.pdf

[21] D. S. Biology, "Deep Space Biology." [Online]. Available: https://deepspace.bio/

[22] LunCo, "Open Source Software for Lunar Colony Engineering." [Online]. Available: https://lunco.space/

[23] I. Otero-Muras and P. Carbonell, "Automated engineering of synthetic metabolic pathways for efficient biomanufacturing," *Metabolic Engineering*, vol. 63, pp. 61–80, Jan. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1096717620301828

[24] M. B. Holowko, E. K. Frow, J. C. Reid, M. Rourke, and C. E. Vickers, "Building a biofoundry," *Synthetic Biology*, vol. 6, no. 1, p. ysaa026, Feb. 2021. [Online]. Available: https://academic.oup.com/synbio/article/doi/10.1093/synbio/ysaa026/6039187

[25] B. A. Bartley, J. Beal, J. R. Karr, and E. A. Strychalski, "Organizing genome engineering for the gigabase scale," *Nature Communications*, vol. 11, no. 1, p. 689, Feb. 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-14314-z

[26] L. M. Sanders, R. T. Scott, J. H. Yang, A. A. Qutub, H. Garcia Martin, D. C. Berrios, J. J. A. Hastings, J. Rask, G. Mackintosh, A. L. Hoarfrost, S. Chalk, J. Kalantari, K. Khezeli, E. L. Antonsen, J. Babdor, R. Barker, S. E. Baranzini, A. Beheshti, G. M. Delgado-Aparicio, B. S. Glicksberg, C. S. Greene, M. Haendel, A. A. Hamid, P. Heller, D. Jamieson, K. J. Jarvis, S. V. Komarova, M. Komorowski, P. Kothiyal, A. Mahabal, U. Manor, C. E. Mason, M. Matar, G. I. Mias, J. Miller, J. G. Myers, C. Nelson, J. Oribello, S.-m. Park, P. Parsons-Wingerter, R. K. Prabhu, R. J. Reynolds, A. Saravia-Butler, S. Saria, A. Sawyer, N. K. Singh, M. Snyder, F. Soboczenski, K. Soman, C. A. Theriot, D. Van Valen, K. Venkateswaran, L. Warren, L. Worthey, M. Zitnik, and S. V. Costes, "Biological research and self-driving labs in

deep space supported by artificial intelligence," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 208–219, Mar. 2023. [Online]. Available: https://www.nature.com/articles/s42256-023-00618-4

[27] N. Gurdo, D. C. Volke, D. McCloskey, and P. I. Nikel, "Automating the design-build-test-learn cycle towards next-generation bacterial cell factories," *New Biotechnology*, vol. 74, pp. 1–15, May 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S187167842300002X

[28] NASA, "About Analog Missions." [Online]. Available: https://www.nasa.gov/analogs/what-are-analog-missions

[29] R. Amils, D. Fernández-Remolar, and the IPBSL Team, "Río Tinto: A Geochemical and Mineralogical Terrestrial Analogue of Mars," *Life*, vol. 4, no. 3, pp. 511–534, Sep. 2014. [Online]. Available: http://www.mdpi.com/2075-1729/4/3/511

[30] S. M. Abramov, D. Straub, J. Tejada, L. Grimm, F. Schädler, A. Bulaev, H. Thorwarth, R. Amils, A. Kappler, and S. Kleindienst, "Biogeochemical Niches of Fe-Cycling Communities Influencing Heavy Metal Transport along the Rio Tinto, Spain," *Applied and Environmental Microbiology*, vol. 88, no. 4, pp. e02 290–21, Feb. 2022. [Online]. Available: https://journals.asm.org/doi/10.1128/aem.02290-21

[31] S. S. Johnson, L. Amaral-Zettler, B. Haas, R. Amils, C. E. Carr, and G. Ruvkum, "Metagenomic Sequencing of the Rio Tinto," *Astrobiology Science Conference 2010*, Apr. 2010.

[32] A. Daddaoua, C. Álvarez, M. Oggerin, N. Rodriguez, E. Duque, R. Amils, J. Armengaud, A. Segura, and J. L. Ramos, "Rio Tinto as a niche for acidophilus enzymes of industrial relevance," *Microbial Biotechnology*, pp. 1751–7915.14 192, Feb. 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/1751-7915.14192

[33] A. Navid, Ed., *Microbial Systems Biology: Methods and Protocols*, ser. Methods in Molecular Biology. New York, NY: Springer US, 2022, vol. 2349. [Online]. Available: https://link.springer.com/10.1007/978-1-0716-1585-0

[34] M. Ponce-de-León, F. Montero, and J. Peretó, "Solving gap metabolites and blocked reactions in genome-scale models: Application to the metabolic network of Blattabacterium cuenoti," *BMC Systems*

Biology, vol. 7, no. 1, p. 114, Dec. 2013. [Online]. Available: https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-114

[35] C. Diener, S. M. Gibbons, and O. Resendis-Antonio, "MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota," *mSystems*, vol. 5, no. 1, pp. e00 606–19, Feb. 2020. [Online]. Available: https://journals.asm.org/doi/10.1128/mSystems.00606-19

[36] A. Zelezniak, S. Andrejev, O. Ponomarova, D. R. Mende, P. Bork, and K. R. Patil, "Metabolic dependencies drive species co-occurrence in diverse microbial communities," *Proceedings of the National Academy of Sciences*, vol. 112, no. 20, pp. 6449–6454, May 2015. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1421834112

[37] F. Zorrilla, F. Buric, K. R. Patil, and A. Zelezniak, "metaGEM: Reconstruction of genome scale metabolic models directly from metagenomes," *Nucleic Acids Research*, vol. 49, no. 21, pp. e126–e126, Dec. 2021. [Online]. Available: https://academic.oup.com/nar/article/49/21/e126/6382386

[38] B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon, "RetroPath2.0: A retrosynthesis workflow for metabolic engineers," *Metabolic Engineering*, vol. 45, pp. 158–170, Jan. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1096717617301337

[39] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Research*, vol. 41, no. D1, pp. D36–D42, Nov. 2012. [Online]. Available: http://academic.oup.com/nar/article/41/D1/D36/1068219/GenBank

[40] S. M. D. Seaver, F. Liu, Q. Zhang, J. Jeffryes, J. P. Faria, J. N. Edirisinghe, M. Mundy, N. Chia, E. Noor, M. E. Beber, A. A. Best, M. DeJongh, J. A. Kimbrel, P. D'haeseleer, S. R. McCorkle, J. R. Bolton, E. Pearson, S. Canon, E. M. Wood-Charlson, R. W. Cottingham, A. P. Arkin, and C. S. Henry, "The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes," *Nucleic Acids Research*,

vol. 49, no. D1, pp. D575–D588, Jan. 2021. [Online]. Available: https://academic.oup.com/nar/article/49/D1/D575/5912569

[41] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko, "The RAST Server: Rapid Annotations using Subsystems Technology," *BMC Genomics*, vol. 9, no. 1, p. 75, Dec. 2008. [Online]. Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-75

[42] G. van Rossum, "Python tutorial," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.

[43] P. Jupyter, "Jupyter Notebook." [Online]. Available: https://jupyter.org/

[44] A. Hunt and D. Thomas, *The Pragmatic Programmer: From Journeyman to Master*. Reading, Mass: Addison-Wesley, 2000.

[45] T. Dalzell, Ed., *The Routledge Dictionary of Modern American Slang and Unconventional English*, second edition ed. London : New York: Routledge, 2018.

[46] R. Jeffries, A. Anderson, and C. Hendrickson, *Extreme Programming Installed*, ser. The XP Series. Boston: Addison-Wesley, 2001.

[47] R. C. Martin, "Design Principles and Design Patterns," 2000. [Online]. Available: https://wnmurphy.com/assets/pdf/Robert_C._Martin_-_2000_-_Principles_and_Patterns.pdf

[48] International Union of Biochemistry and Molecular Biology and E. C. Webb, Eds., *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992.

[49] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 304–305, Jan. 2000. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.304

[50] M. H. Saier, V. S. Reddy, G. Moreno-Hagelsieb, K. J. Hendargo, Y. Zhang, V. Iddamsetty, K. J. K. Lam, N. Tian, S. Russum, J. Wang, and A. Medrano-Soto, "The Transporter Classification Database (TCDB): 2021 update," *Nucleic Acids Research*, vol. 49, no. D1, pp. D461–D467, Jan. 2021. [Online]. Available: https://academic.oup.com/nar/article/49/D1/D461/5973435

[51] D. A. Cuevas, J. Edirisinghe, C. S. Henry, R. Overbeek, T. G. O'Connell, and R. A. Edwards, "From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model," *Frontiers in Microbiology*, vol. 7, Jun. 2016. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00907/abstract

[52] D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil, "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities," *Nucleic Acids Research*, vol. 46, no. 15, pp. 7542–7553, Sep. 2018. [Online]. Available: https://academic.oup.com/nar/article/46/15/7542/5042022

[53] T. Duigou, M. du Lac, P. Carbonell, and J.-L. Faulon, "RetroRules: A database of reaction rules for engineering biology," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1229–D1235, Jan. 2019. [Online]. Available: https://academic.oup.com/nar/article/47/D1/D1229/5128930

[54] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Data Analysis, Machine Learning and Applications*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 319–326. [Online]. Available: http://link.springer.com/10.1007/978-3-540-78246-9_38

[55] M. Ganter, T. Bernard, S. Moretti, J. Stelling, and M. Pagni, "MetaNetX.org: A website and repository for accessing, analysing and manipulating metabolic networks," *Bioinformatics*, vol. 29, no. 6, pp. 815–816, Mar. 2013. [Online]. Available: https://academic.oup.com/bioinformatics/article/29/6/815/183749

[56] IUPAC, "Project Details: IUPAC - International Chemical Identifier." [Online]. Available: https://web.archive.org/web/20120527162256/

http://www.iupac.org/home/projects/project-db/project-details.html?tx_wfqbe_pi1%5Bproject_nr%5D=2000-025-1-800

[57] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev, "InChI - the worldwide chemical structure identifier standard," *Journal of Cheminformatics*, vol. 5, no. 1, p. 7, Dec. 2013. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-5-7

[58] J. Myung, J. C. A. Flanagan, R. M. Waymouth, and C. S. Criddle, "Expanding the range of polyhydroxyalkanoates synthesized by methanotrophic bacteria through the utilization of omega-hydroxyalkanoate co-substrates," *AMB Express*, vol. 7, no. 1, p. 118, Dec. 2017. [Online]. Available: http://amb-express.springeropen.com/articles/10.1186/s13568-017-0417-y

[59] W. D. Jang, G. B. Kim, and S. Y. Lee, "An interactive metabolic map of bio-based chemicals," *Trends in Biotechnology*, vol. 41, no. 1, pp. 10–14, Jan. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167779922001950

[60] National Center for Biotechnology Information, "PubChem Compound Summary for CID 3541112, 3-Hydroxybutyrate." 2023. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/3-hydroxybutyrate#section=3D-Conformer

[61] J. Lopez-Ibañez, F. Pazos, and M. Chagoyen, "Predicting biological pathways of chemical compounds with a profile-inspired approach," *BMC Bioinformatics*, vol. 22, no. 1, p. 320, Dec. 2021. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04252-y

[62] M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.27

[63] J. Wicker, T. Lorsbach, M. Gütlein, E. Schmid, D. Latino, S. Kramer, and K. Fenner, "enviPath – The environmental contaminant biotransformation pathway resource," *Nucleic Acids Research*, vol. 44, no. D1, pp. D502–D508, Jan. 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1229

[64] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "COBRApy: COnstraints-Based Reconstruction and Analysis for Python," *BMC Systems Biology*, vol. 7, no. 1, p. 74, Dec. 2013. [Online]. Available: https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-74

[65] H. H. Hau and J. A. Gralnick, "Ecology and Biotechnology of the Genus *Shewanella*," *Annual Review of Microbiology*, vol. 61, no. 1, pp. 237–258, Oct. 2007. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev.micro.61.080706.093257

[66] V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, R. J. Xavier, R. Knight, K. Cho, and R. Bonneau, "Structure-based protein function prediction using graph convolutional networks," *Nature Communications*, vol. 12, no. 1, p. 3168, May 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-23303-9

[67] M. Kratochvíl, L. Heirendt, S. E. Wilken, T. Pusa, S. Arreckx, A. Noronha, M. van Aalst, V. P. Satagopam, O. Ebenhöh, R. Schneider, C. Trefois, and W. Gu, "COBREXA.jl: Constraint-based reconstruction and exascale analysis," *Bioinformatics (Oxford, England)*, vol. 38, no. 4, pp. 1171–1172, Nov. 2021. [Online]. Available: https://doi.org/10.1093/bioinformatics/btab782

[68] C. Lieven, M. E. Beber, B. G. Olivier, F. T. Bergmann, M. Ataman, P. Babaei, J. A. Bartell, L. M. Blank, S. Chauhan, K. Correia, C. Diener, A. Dräger, B. E. Ebert, J. N. Edirisinghe, J. P. Faria, A. M. Feist, G. Fengos, R. M. T. Fleming, B. García-Jiménez, V. Hatzimanikatis, W. Van Helvoirt, C. S. Henry, H. Hermjakob, M. J. Herrgård, A. Kaafarani, H. U. Kim, Z. King, S. Klamt, E. Klipp, J. J. Koehorst, M. König, M. Lakshmanan, D.-Y. Lee, S. Y. Lee, S. Lee, N. E. Lewis, F. Liu, H. Ma, D. Machado, R. Mahadevan, P. Maia, A. Mardinoglu, G. L. Medlock, J. M. Monk, J. Nielsen, L. K. Nielsen, J. Nogales, I. Nookaew, B. O. Palsson, J. A. Papin, K. R. Patil, M. Poolman, N. D. Price, O. Resendis-Antonio, A. Richelle, I. Rocha, B. J. Sánchez, P. J. Schaap, R. S. Malik Sheriff, S. Shoaie, N. Sonnenschein, B. Teusink, P. Vilaça, J. O. Vik, J. A. H. Wodke, J. C. Xavier, Q. Yuan, M. Zakhartsev, and C. Zhang, "MEMOTE for standardized genome-scale metabolic model testing,"

Nature Biotechnology, vol. 38, no. 3, pp. 272–276, Mar. 2020. [Online]. Available: https://www.nature.com/articles/s41587-020-0446-y